Data Analytics Skills Evaluation Framework

Midou Nafaa and Frank Mu

CodeSignal
Skills Evaluation Lab

Published August 2021

Abstract - In the Information economy of the 21st century, the most valuable commodity is talent. This is true across all levels of organizations, and especially for data analytics roles that are so integral to organizations' decision making, innovation, and execution. However, the challenge of running an effective data analyst talent acquisition operation is all too familiar: scalability. Organizations must deploy their already scarce resources to widen the recruitment funnel and ultimately broaden and diversify their applicant pool. CodeSignal's Data Analytics Assessment (DAA) framework brings the scaling factor to take your talent acquisition strategy to the next level. With tests created through our scaled DAA platform, you will not have to compromise on quality and business agility to build a data talent competitive advantage.

1. Introduction

With increasingly connected and digitized economies, business competitiveness is becoming more and more dependent on leveraging data insights to inform business decisions across all levels of modern organizations. As such, over the past decade, data analysis has become one of the most in-demand skills in businesses across many industries[1] [1].

While there is a remarkable growth in the number of educational programs dedicated to the *data analysis* field, the reality is that successful data analysts can have different educational backgrounds (e.g., mathematics, statistics, economics, finance, marketing, business, or computer science). Therefore, it is critical for any competitive recruitment process for data analysts to intentionally be broad and inclusive. The challenge is scalability. Time and resources committed to sourcing candidates will always limit organizations' ambitions to ramp-up hiring pace and its aspirations to diversify the talent pool.

The status quo in today's technical recruiting process is to use a resume/CV as a proxy for skills, which tends to lead to biased and inefficient recruiting and evaluation practices. Recognizing the limits of CV-based candidate sourcing, many organizations have moved towards creating their own assessments, which are often used as the first hurdle in the hiring process. However, because many organizations do not have expertise in designing assessments for hiring, this approach may lead to flawed tests [2] that may fall short of complying with EEOC guide-

[1]According to the U.S. Bureau of Labor Statistics, management analyst jobs will see a 1% growth from 2019 to 2029 (much faster than average).

lines for employment tests[2] [3]. Furthermore, organizations will have to take on the significant burden of actively managing plagiarism risk to maintain the integrity and effectiveness of their custom assessments, as resources spent on creating assessment content will be wasted if test-takers leak the test content to websites like Glassdoor, StackOverflow, etc.

Effective recruitment operations should still include candidate interviews to assess advanced skills, intangible attributes, and role/culture fit. However, we argue that a well-designed top-of-funnel assessment delivered over an online platform can significantly enhance existing hiring operations by ***(i)*** scaling where/how candidates can be sourced without compromising on the quality and diversity of candidates; ***(ii)*** ensuring consistency and fairness when assessing for fundamental data analysis skills, and ***(iii)*** providing effective signals about candidates' skills to streamline the recruitment funnel and reduce time spent interviewing unqualified candidates.

CodeSignal's **Data Analytics Assessment (DAA) framework** was specifically designed to effectively evaluate some of the fundamental data analysis skills valued by hiring managers across a variety of industries. The assessment framework enables recruitment teams to dramatically increase their candidate sourcing footprint by allowing them to generate standardized tests that are comprehensive, consistent, EEOC-compliant, and easily scalable to a large pool of questions/tasks. Additionally, this framework allows simple management of plagiarism risk, as content leaked to sites like Glassdoor or StackOverflow can easily be replaced with other content generated from the framework.

2. Building a Competitive Advantage with CodeSignal's DAA Framework

At the highest level, running an effective recruitment operation will always come down to balancing a number of tradeoffs, such as targeting a wide talent pool, building process agility, and rationalizing internal resources allocated to hiring. In the following, we discuss how online assessment can add concrete business value:

1) **Tapping into the Widest Talent Pool:** A widely accepted benefit of broader candidate sourcing resides in maximizing the chances of hiring the most qualified candidates. However, an equally important benefit is talent diversification, which has been found to produce better business outcomes in terms of decision making, team dynamics, and financial ROI [4]. Further, workplace diversity is today an expectation by employees, shareholders, and customers. Using the DAA framework can deliver the much needed scalability factor (wider top of the funnel) so your pursuit of talent diversity doesn't come with prohibitive resource requirements.

[2] The U.S. Equal Employment Opportunity Commission (EEOC) is responsible for enforcing federal laws that make it illegal to discriminate against a job applicant.

2) **Optimizing the Interview Funnel:** A large-scale study [5] of recruitment funnels across industries highlights a surprising and counterintuitive finding: the effectiveness of the lower-cost upstream recruitment process is surprisingly comparable to the effectiveness of the more costly downstream recruitment process. In other words, interview-to-offer ratio (~20%) should be much higher than the applicants-to-interview ratio (~15%). The DAA framework can help address this cost-value process imbalance by enforcing rigorous standards to qualify applicants for the interview stages. This will ultimately increase the interview-to-offer-ratio, and thus reduce cost per hire.

3. **Partnership between Hiring Manager (HM) and Recruiter:** Ongoing alignment between the HM and recruiter is critical for the health and effectiveness of any recruitment process. This is usually achieved by continuously communicating on candidate-role fit, which is when the coordination overhead can start getting in the way of business agility. The DAA framework can be used to ensure consistency in core knowledge and skill requirements, thus reducing the burden of frequent back-and-forth communications between HMs and recruiters.

Clearly, producing a higher-quality interviewee pool will improve efficiencies, minimize the required interviewing resources, while also improving your overall business agility (time-to-hire).

CodeSignal's DAA framework can dramatically scale an organization's ability to both widen the data talent pool (quality), all while maintaining high standards for qualifying applicants to the interview stage (cost).

3. The Framework

It is important to note that the DAA framework was intentionally designed to focus on assessing core data analytic skills, as such, skills are transferable across most industries[3]. This will help avoid unintentional biases towards certain industries, data job families, or data tools. Moreover, this will also avoid unintentionally assessing test-takers on "niche" data skills that should be developed once they are working in the role

3.1 Topics

The DAA framework is based on the general business wisdom around skills and competencies that make for a successful data analyst[4].

At the highest level, successful data analysts are able to independently perform tasks associated with the entire lifecycle of answering business questions with data, including:

1. **Analytic thinking:** One of the

[3] https://www.coursera.org/articles/what-does-a-data-analyst-do-a-career-guide

[4] "*Advancing Your Analytics Career*", Northeastern University, 2021.

most important competencies that determines the ability to connect the dots and process information, while anchoring on the business context. This is a good predictor of the test-taker's ability to plan and execute a complex analysis.

2. **Manipulating data with purpose:** Beyond the technical data skills (Excel, SQL, Python, etc.), a key competency for data analysts across the board is the ability to purposefully plan and execute a series of data manipulations in an efficient manner. This will enable analysts to conduct deeper analyses that may be idiosyncratic to the specific situation.

3. **Effectively communicating business insights:** This is an important skill that's critical for turning analyses and results into high-quality insights that can be used in business decisions. This is about summarizing the data analysis output to effectively communicate the key actionable insights (with visualization and business context).

While some of these skills and competencies are better assessed in one-on-one settings, it's still worth summarizing what we believe are the fundamental competencies that predict success in the workplace.

3.2 Test Structure

Based on the topics discussed in the previous section 3.1, each test has **3 groups of tasks**. Table 1 summarizes the overall structure of the DAA test. In total, there are **15 tasks**: **1 database manipulation task**, which requires 10-20 lines of SQL code, and **14 quiz tasks** grouped into 6 categories. Table 2 summarizes the quiz tasks. The maximum allowed completion time for the test is **70 minutes**; however, candidates are not necessarily expected to complete all tasks within this time. Part of assessing candidates' skill levels is to see how far they can progress in the DAA within the given time frame. Given the emphasis on core data analytics skills and competencies, tests based on the DAA framework should focus on assessing analytical thinking and data manipulation skills/competencies while remaining agnostic to different data tools or coding languages (e.g., Google Sheets, Microsoft Excel, R, or Python). We recommend evaluating the more nuanced communications skills and tool or language specific skills during the interview phase of the hiring process.

Table 2 summarizes what the DAA online test taker would be exposed to in terms of multiple-choice questions and what data skill category they fall under.

3.3 Creating tasks for DAA

Group 1 – Basic Analytics (Quiz Tasks)

This group consists of **6 quiz tasks** covering 3 topics: data processing, basic statistics, and basic data aggregation. These tasks all revolve around a common data scenario with only 1 dataset (within 1 input table). The expected total time for solving all the tasks in this group is between **10 and 15 minutes**. Task parameters are described below and a task example (3.3.1) can be found in the appendix of this document.

Expected Knowledge

- Simple data cleaning/processing, such as ensuring each row of data has appropriate identifiers
- Selecting, filtering, and sorting spreadsheet/tabular data
- String and numeric data types
- Aggregating/summarizing large data tables into smaller summary tables based on business context and questions
- Basic statistical concepts, such as min, max, average, median, percentages/ proportions, ratios

Can Include

- Tasks that implicitly require data aggregation/summarizing, such as reporting the group of data points with highest sum or average
- Tasks that require computations using the entire dataset, such as computing percentages across groups of data points
- Tasks that require simple comparison across multiple groups of data points, such as comparing sums/averages, but not requiring statistical analyses and inference
- All tasks should revolve around one scenario with a single dataset/input table of medium size (<500 rows) that's easily accessible via most data analytics tools

Should Exclude

- Tasks that explicitly instruct test-takers to perform a certain operation
- Tasks requiring data merging/joins, such as working with multiple datasets/input tables
- Tasks that do not require manipulating the dataset, such as locating a specific row, column, or value
- Tasks that require knowledge of statistical analyses and inference
- Tasks that require complex data transformations

Group 2 – Data Manipulation/Databases (SQL Task)

This group consists of **1 code writing task** in Query Language (standard SQL). The expected total time for solving this SQL task is between **15 and 25 minutes**. Task parameters are described below and a task example (3.3.2) can be found in the appendix of this document.

Expected Knowledge

- Basic SQL operations and functions, including:
    - Selecting, filtering, sorting
    - Aggregate and count functions
    - If and case functions
    - String functions
    - Subqueries/inner-queries
    - Joins: inner, left, right, outer
    - Window functions
    - Window aggregate functions

Can Include

- Tasks which require writing SQL code to conduct basic SQL operations described above
- Tasks like example 3.3.2 (see ap-

pendix), which assesses a combination of SQL selecting, filtering, sorting, joins, etc.

Should Exclude

- Complex queries which require deeper knowledge of SQL operations and functions not discussed above
- Tasks which require dealing with non-tabular data – JSON, XML, etc.

Group 3 – Advanced Analytics (Quiz Tasks)

This group consists of **8 quiz tasks** covering 3 topics – data merging/joins (outside of SQL context), conditional data aggregation, and basic statistical inference. These tasks all revolve around a common data scenario with multiple datasets (e.g, 3-5 input tables). The expected total time for solving all the tasks of this group is between **25 and 30 minutes**. Task parameters are described below and a task example (3.3.3) can be found in the appendix of this document.

Expected Knowledge

- Advanced data cleaning/processing, such as ensuring each row of data has appropriate identifiers and handling missing values
- Selecting, merging/joining, filtering, and sorting spreadsheet/tabular data across several input tables
- String and numeric data types
- Aggregating and summarizing large data tables into smaller summary tables based on business context and questions
- Basic statistical concepts such as min, max, average, median, percentages/proportions, ratios
- Basic statistical inference (e.g., t-tests and p-values)

Can Include

- Tasks that require merging data from multiple tables together, then aggregating/summarizing the combined data to extract insights
- Tasks that require computations using the multiple data tables
- Tasks that require comparing across multiple groups of data points using appropriate statistical analyses and inferences
- All tasks should revolve around one scenario with multiple dataset/input tables of medium-large size (500-2000 rows) that are easily accessible via most data analytics tools

Should Exclude

- Tasks that explicitly instruct test-takers to perform a certain operation, such as computing average for a specific group
- Tasks that do not require manipulating the datasets, such as locating a specific row, column, or value
- Anything that requires complex data transformations
- Anything that requires complex statistical analyses or machine learning algorithms

Fundamental DA Competence	Skill Assessment Description	Skill Assessment Goals	Time Est.(% of score)
Basic Analytical Thinking	Simple data manipulations to answer business questions	Assessing basic abilities to work with spreadsheet-s/tabular data to process and summarize data	10-15 mins for 6 quiz tasks
Data Manipulation (Database Queries)	SQL code writing, with low-moderate difficulty	Assessing basic SQL cod-ing skills, ability to work with databases, and ability to summarize large data ta-bles to present key informa-tion	15-25 min for 1 SQL task
Advanced Analytical Thinking + Data Manipulation Across Mul-tiple Tables	Moderately difficult data manipulations involving multiple datasets, and basic statistical inference (understanding p-values from t-tests), to answer business questions	- Assessing moderate-advanced abilities to work with multiple spreadsheets/tabular datasets simultaneously, and ability to anchor on the business context to guide the analysis - Assessing ability to draw valid conclusions from data and think through basic business statistics concepts (distribution, standard deviation, etc.)	25-30mins for 8 quiz tasks

Table 1: High-Level Structure of the Data Analytics Assessment Framework

Quiz Groups		Quiz tasks count	Total tasks in group
Basic Analytics	**Data processing**	**2**	**6**
	Basic statistics	**2**	
	Basic aggregation	**2**	
Advanced Analytics	**Data merging/joins**	**4**	**8**
	Conditional aggregation	**2**	
	Statistical Inference	**2**	

Table 2: Summary of the Set of DAA Quiz Groups

3.4 Test Scoring

The scoring system for tests that can be generated by the DAA framework includes three (3) important properties: **balanced scoring**, **interpretability**, and **deeper insights into strengths**.

1. **Balanced scoring:** The framework and scoring were designed to create balanced tests that appropriately considers all data analyst skills needed throughout the lifecycle of answering business questions with data. We intentionally include multiple tasks related to analytical thinking because this is a foundational data intuition skill that underpins many analyst competencies which can take a long time to develop.

2. **Interpretability:** The result of the test is an overall data analytics ability score based on the test-taker's solved

tasks and speed of completion. This score is a number ranging from 600-850, and can be interpreted in a similar manner as FICO credit scores. Figure 1 illustrates some of the most common test outcomes arranged by ability levels (expected scores). For simplicity, in Table 3 we will assume that a test-taker either solves all the tasks from a quiz group, or does not solve any tasks. Descriptions of the test-taker's ability are provided for each level within the table.

3. **Deeper insights into strengths:** In addition to the overall score, the test also produces a detailed report for each test-taker, which includes itemized scores and completion times for each task. This detailed score report is particularly useful to help recruiters and hiring managers do deep dives into the candidate's skills and competencies, which allow them to focus on identified gaps at later stages in the hiring process.

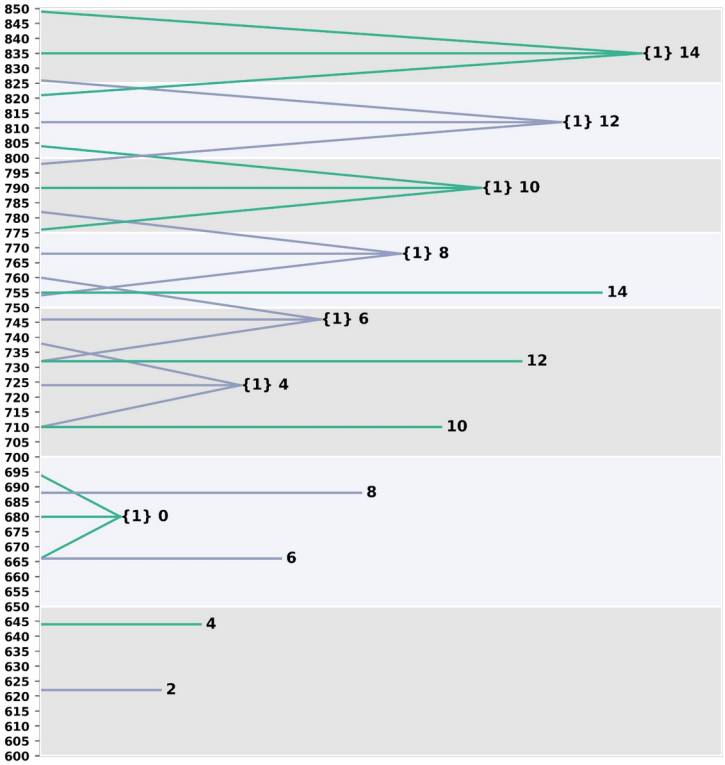

Figure 1. The score ranges based on which tasks a test-taker has solved.

Expected Score	SQL	Quiz Groups	Description
665	No	1st	The test-taker can answer business questions by doing simple data aggregation and comparing descriptive statistics.
680	Yes	0	The test-taker can write advanced SQL queries to summarize large datasets.
746	Yes	1st	The test-taker can answer business questions by writing SQL queries or using other tools to aggregate data and extract insights.
768	Yes	2nd	The test-taker can answer business questions by writing SQL queries and using other tools to merge data, compare descriptive statistics, and make basic statistical inferences.
835	Yes	1st, 2nd	The test-taker is an experienced data analyst.

Table 3: High-Level Description of the Range of Scores of Candidates

3.5 Results

The DAA has been reviewed and used by subject matter experts with data analytic backgrounds and piloted within the selection process for various data analytic positions, across different business areas. Figures 2 and 3 are the distributions of scores from an initial sample of 147 test-takers.

As seen in Figures 2 and 3, this initial sample offers support for the DAA's ability to distin-

guish between different levels of DA skills. As this is a relatively small pilot sample, however, we do expect distributions to vary depending on the sourcing strategy, business area, and skill levels being assessed within various selection processes. The intent is to release an updated version of these figures once more data become available.

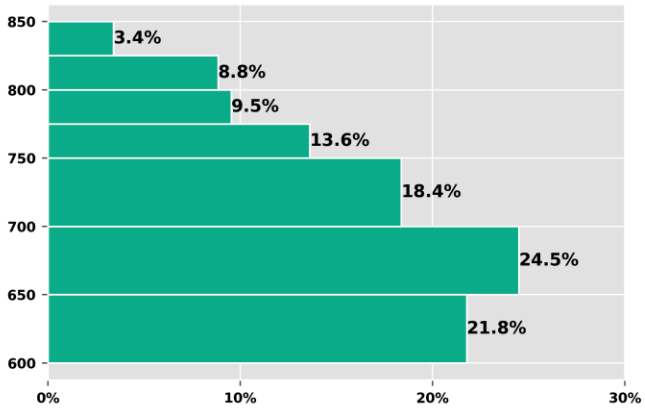

Figure 2: The percentage distribution of scores for a sample of 147 test-takers.

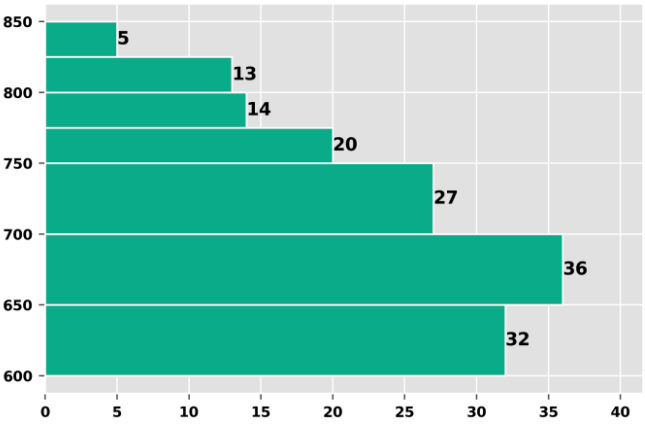

Figure 3: The histogram of scores for a sample of 147 test-takers.

4. Framework Development

The design of the DAA framework was informed by an extensive review of literature from the following sources:

1. We've examined data analytics topics covered in University courses (including online courses via Coursera and edX), with a particular focus on the topics covered by many courses (with implied importance). In particular, we've compared some of the most popular and highly rated online courses by Johns Hopkins University [6], University of Pennsylvania [7], Harvard University [8], Massachusetts Institute of Technology [9], and Columbia University [10].

2. We've surveyed four data analysts with 3 to 8 years of industry experience about interview questions they received (as interviewee) and/or they administered (as interviewer) as well as about the topics they think are critical for data analysts.

3. We've compared professional data analyst certifications (such as Microsoft Certified: Data Analyst Associate [11], Google Data Analytics Professional Certificate [12], IBM Data Analyst Professional Certificate [13]). Although the exams for these certificates focus on a variety of topics and differ from one certification to another, the most overlapped and foundational topics are included in this assessment framework.

4. Based on our research using these relevant materials and experience running data analytics functions, we identified the topics discussed in section 3.1 as the most important foundational concepts in data analytics. Then, we narrowed these topics to ones which can be automatically evaluated at scale via online assessments in a fair, consistent, and valid manner, and identified topics listed in Table 2 as the best framework for scalably developing pre-hire assessments.

5. Why the CodeSignal Platform?

Besides leveraging CodeSignal's platform and experience in the online assessment space, the DAA framework was specifically designed to be a consistent, fair, and comprehensive way to assess foundational data analytics skills and competencies. Advantages of tests developed from this framework include:

1. **Effectiveness and integrity of the online tests:** The risk of plagiarism and content leaks are significantly reduced with CodeSignal's ability to generate fully randomized data sets for each group of tasks in the test. Additionally, this framework approach allows for alternating between different scenarios that vary fundamental elements of the data problems without affecting the consistency of the overall assessment.

2. **Data analyst recruitment operation insights:** As the leading player in online technical assessments across industries, CodeSignal is uniquely positioned to share unparalleled insights and benchmarking on the effectiveness of hiring operations for data analytics roles. Further, linking assessment data with data from interview outcomes and workplace performance will enable you to unlock critical business insights to build a competitive advantage around data-related talent acquisition.

3. **Evolution and continued market relevance:** CodeSignal is committed to maintaining the DAA framework to ensure that the assessment is inclusive, comprehensive, and focused on the core skills that predict candidate success in the workplace.

6. Considerations and Recommendations

1. **Visualization is best assessed during interviews:** Clearly, effectively communicating data insights (including visualizations) is a critical skill that's often the difference between a good and bad business decision. However, it is challenging to evaluate data communications skills through online tests at scale without introducing biases. The most tangible data communication competence resides in how a data analyst succinctly distills data insights and presents it a structured and actionable way. This is better assessed in the conversational environment that interviews offer.

2. **Making the most of interview time:** As discussed earlier, the value of the DAA framework extends beyond scaling the upstream part of the hiring operation (sourcing). The assessment can also help streamline the most time-consuming and costly part of recruiting: interviews. First, the detailed online assessment report can be leveraged to improve the focus for interviews at later stages of the hiring process. Second, aggregated data on test performances can be analyzed to glean valuable insights to refine interview content, which help to improve the interview-to-hire ratio.

3. **Analytical thinking is a requirement for any modern business role:** While our Data Analyst Assessment was designed to scale

data analysts hiring, we believe a large portion of the test is relevant for improving the hiring effectiveness for any entry-level business operation and/or strategy role. With the increasingly data-driven economy of today, the analytical thinking skills assessed by this testing framework are necessary for higher-quality decision making, the lifeblood of modern organizations.

References

[1] U.S. Bureau of Labor Statistics. Occupational Outlook Handbook - Management Analyst. [Online]. Available: https://www.bls.gov/ooh/business-and-financial/management-analysts.htm#tab-6

[2] A. Sahakyan and T. Sloyan, "General coding assessment framework," CodeSignal Research, 2019. [Online]. Available: https://codesignal.com/general-coding-assessment-framework/

[3] U.S. Equal Employment Opportunity Commission. Employment Tests and Selection Procedures. [Online]. Available: https://www.eeoc.gov/laws/guidance/employment-tests-and-selection-procedures

[4] P. Gompers, and S. Kovvali, "The Other Diversity Dividend", Harvard Business Review, Aug 2018

[5] R. Shetelboim, W. Hsu, A. Van Nuys, "2017 Recruiting Funnel Benchmark: analysis and actional tip to improve recruiting performance", Jobvite.

[6] J. Leek et al. Getting and Cleaning Data. [Online]. Available: https://www.coursera.org/learn/data-cleaning

[7] E. Bradlow et al. Business Analytics Specialization. [Online]. Available: https://www.coursera.org/specializations/business-analytics

[8] R. Irizarry. Data Science: Wrangling. [Online]. Available: https://online-learning.harvard.edu/course/data-science-wrangling

[9] E. Duflo and S. F. Ellison. Data Analysis for Social Scientists. [Online]. Available: https://www.edx.org/course/data-analysis-for-social-scientists

[10] A. Gelman et al. Statistical Thinking for Data Science and Analytics. [Online]. Available: https://www.edx.org/course/statistical-thinking-for-data-science-and-analytic

[11] Microsoft. Microsoft Certified: Data Analyst Associate. [Online]. Available: https://docs.microsoft.com/en-us/learn/certifications/data-analyst-associate/

[12] Google Career Certificates on Coursera. Google Data Analytics Professional Certificate. [Online]. Available: https://www.coursera.org/professional-certificates/google-data-analytics

[13] R. Ahuja et al. IBM Data Analyst Professional Certificate. [Online]. Available: https://www.coursera.org/professional-certificates/ibm-data-analyst

Author Bios

Midou Nafaa is a Director of Analytics at Google where he leads the customer success analytics group dedicated to identifying, understanding and addressing customer pain points within and across flagship Google Products. Prior to joining Google, Midou had various start-up experiences as co-founder and Product lead in the computer networking space. He received a PhD degree in Computer Science from University of Versailles, France in 2005 and an MBA from Boston University, USA in 2012.

Frank Mu is a Senior Assessment Research Manager at CodeSignal where he plays a critical role in ensuring scientific rigor behind CodeSignal assessments in terms of how they are developed, monitored, and used. He received his PhD in Industrial-Organizational Psychology from the University of Waterloo and is an active member of SIOP, serving as a volunteer on the Membership Analytics Subcommittee.

Appendix

Example 3.3.1

Scenario

A small company develops a platform for video streaming. The company wants to perform A/B testing of their ads display algorithms. The company has two versions of ads display algorithms: version A and version B.

The experiment took place for one month and the aggregated performance logs are gathered in the dataset below. The data is structured by day and time range (the number of time ranges is fixed for each day). However, due to limitations of systems logging, the overall experiment data is reported in an aggregated form for both experiments (A + B), and it is also reported separately for experiment B; unfortunately, data for experiment A is not reported separately.

Using the dataset below, please, answer the following questions:
[Test takers are asked to download the dataset (data2.csv)]

Data Explanation:

In the first row: 2020-06-01,00:00-05:59,409671,26195,22,299059,24164,13

On June 1st 2020 during the time range 00:00-05:59, in total (during experiments for both versions A and B) users have spent 409671 minutes watching the content and ads, 26195 minutes watching only ads, and at the same time users have followed 22 ads links. For experiment of version B, users have spent 299059 minutes watching the content and ads, 24164 minutes watching only ads, and at the same time followed 13 links in total for this experiment.

Sample Tasks:

1. For a given time range, let's define *ad_time_ratio* as a ratio of time spent for watching ads over the total time spent on the platform. Which date has the highest average *ad_time_ratio* across all time intervals?
   - To solve:
     - Make sure that all time interval values across all rows are labelled correctly
     - Create column: *ad_time_ratio = total_ads_watched_in_mins/total_user_time_spent_in_mins*
     - Aggregate data to time interval level by grouping all rows with the same time interval value (e.g., 00:00-05:59, 06:00-11:59, etc.) together, then computing the average of *ad_time_ratio* for each group
     - Sort the aggregated data in descending order and report the time interval with the highest average *ad_time_ratio*
2. Which time range has the smallest percentage of watched ads over all the ads watched?
   - To solve:
     - Make sure that all time range values across all rows are labelled correctly
     - Compute total ads watched by summing *total_ads_watched_in_mins* across entire dataset

- Aggregate data to time range level by grouping all rows with the same time range value (e.g., 00:00-05:59, 06:00-11:59, etc.) together, then computing the sum of t*otal_ads_watched_in_mins* for each group
- Divide the sums from the aggregated data by the total ads watched value to get percentage of watched ads for each time range
- Sort the final list of values above in ascending order, then report the time range with the smallest percentage of watched ads

Example 3.3.2

Scenario

The toy factories have already finished producing all the presents for the holiday shopping season, but before workers can start delivering them they need to be properly packaged.

All produced gifts and available packages are stored in two tables called **gifts** and **packages** respectively, that have the following structures:

- **gifts**:
  - *id*: unique gift id;
  - *gift_name*: the name of the gift;
  - *length*: gift length;
  - *width*: gift width;
  - *height*: gift height;
- **packages**:
  - *package_type*: package type;
  - *length*: the length of the package;
  - *width*: the width of the package;
  - *height*: the height of the package.

A gift fits in a package if its length, width and height are equal to or less than length, width and height of the package respectively. Note that the presents can't be rotated, since some of them are very fragile.

There is not much space on the delivery trucks, so each gift is put in the smallest package in which it fits. One package is considered to be smaller than the other if its volume is smaller than the volume of the other package. Note, that one package can't hold more than one gift.

Given the tables gifts and packages, compose the resulting table with two columns: package_type and number. The first column should contain the package_type of the package, and the second column should contain the number of the packages with such package_type that will be used for packaging holiday season gifts in the manner described above. If a package of some type wasn't used at all, it shouldn't be included in the result.

The result should be sorted by the package_type column in ascending order.

It is guaranteed that each gift fits some package and that there are no package types with the same volume.

For the following tables **gifts**:

gifts

package_type	length	width	height
big	4	4	4
extra	5	5	5
medium	2	2	2
small	1	1	1
special	4	3	1

and **packages**:

packages

the output should be:

package_type	number
big	2
medium	1
small	1
special	1

output

id	gift_name	length	width	height
1	Water gun	3	1	1
2	Video game	1	1	1
3	Toy car	4	2	2
4	Toy car	4	2	2
5	Toy gun	2	1	1

Note that there is no row for extra package type in the output, since it won't be used.

Sample Solution (in MySQL):

```
CREATE PROCEDURE giftPackaging()
  select
       (select package_type from packages
        where g.length <= length &&
              g.width <= width &&
              g.height <= height
        order by length * width * height
        limit 1) package_type,
        count(1) number
   from gifts g
   group by 1
   order by 1
```

Example 3.3.2

Scenario

You are analyzing data for a company specializing in video advertisements. The data is scattered across 4 .csv files:

ads.csv contains the names of the advertisements and corresponding IDs for the full and shortened versions of the video files.

ads.csv

ad_id,ad_name,long_version_video_id,short_version_video_id

0,"chats-conditionate",243,392

1,"isomerizing-louma",1694,1455

...

videos.csv contains information about the video files: the duration of each video (in seconds), and the path to the video file in the filesystem.

videos.csv

video_id,path,duration

1000,"/root/bucket5/halfwitted-steganographs.mp4",59

1001,"/root/bucket2/plottier-tortoises.mp4",54

...

platforms.csv contains information about different platforms, which broadcast the advertisements. Note that some information in this table can be missing (null).

platforms.csv

platform_id,contact_mail,website

0,"kmiller@yahoo.com","comcast.net"

1,null,"japanpost.jp"

2,"clkao@gmail.com",null

...

ads_statistics.csv contains broadcasting metrics. Note that the total time is in seconds, and the price is in dollars.

ads_statistics.csv

platform_id,video_id,watch_count,total_time_watched,price_per_watch

102,1770,328483,18395048,0.32

1,1154,986722,38482158,0.22

...

Sample Tasks:

1. Find the number of rows in ads_statistics.csv with unknown platform_id. Note that platform_id is unknown if it's not in the platforms.csv file.

- To solve:
  - Ensure that id columns in both datasets all have unique values
  - Merge platforms.csv to ads_statistics.csv via a left join, with ads_statistics.csv as the referent (left) table and platform_id as the key id column
  - If successful, in the merged dataset, columns (i.e., contact_mail and website) should contain null/missing values for rows if platform_id exists in ads_statistics.csv but not in platforms.csv
  - Filter out all such rows/cases where columns contact_mail and website contain null/missing values
  - Count and report the number of remaining roles

2. Imagine that a business leader in the company wants to implement a "low performing platform" program which will prioritize investing in platforms with average watch_count that is statistically significantly lower than a benchmark platform. Initially, platform_id=250 was selected as the benchmark platform. The platform managers for platform_id=178, platform_id=143, and platform_id=169 are all eager to enroll their platforms into the program. Which of these platforms have a significantly lower average watch_count compared to platform_id=250? Please assume a significance level of $\alpha = 0.05$.

- To solve:
  - Identify relevant dataset - ads_statistics.csv
  - Ensure that the id column has unique values across all rows
  - Conduct t-tests comparing average watch_count of all other platforms against average watch_count of platform_id=250
  - From the results of t-tests, determine which comparisons meet the specified significance level to identify the relevant platforms for this context