Data Analytics
Skills Evaluation Framework

Technical Brief

CodeSignal

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.

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, inclusive, and scalable. 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.

This paper describes a framework for developing simulation-based evaluations that accurately capture high quality signals about the technical knowledge and core skills required for Data Analytics roles at scale. Framework-based evaluations are expertly designed and highly structured, allowing data and talent teams to efficiently scale their hiring process and make effective hiring decisions, while providing a fair and engaging experience for candidates.

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

1. **Analytical thinking:** One of the most important skills that determines the analyst's ability to connect the dots and process data, while anchoring on the business context.
2. **Manipulating data with purpose:** Beyond the knowing how to use data tools (Excel, SQL, Python, etc.), data analysts must be able to purposefully plan and execute a series of data manipulations in an efficient manner.
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 Framework, developed based on researching data analyst jobs and consultation with subject matter experts, is designed to assess the key knowledge and skills that are commonly required for data analyst roles across a wide variety of organizations and industries.

Framework Specifications

The Data Analytics Skills Evaluation framework is designed to closely simulate the fundamental knowledge and skills a candidate would be expected to have within data analytics roles. The framework can be utilized to create evaluations that span different delivery methods, such as pre-screen assessments or technical interviews, while providing objective signals by automatically generating scores to quantify candidates' skill levels.

Evaluations based on this framework consist of three modules that target a breadth of data analytics topics:

- Data cleaning & manipulation
- Descriptive statistics
- Databases & SQL
- Extracting insights

Candidates will be demonstrating key analytics knowledge and skills by effectively solving questions within the modules. To balance the depth and breadth of content and candidate experience, **the evaluation time for this framework is 70 minutes**. Possible scores range from 200 to 600.

Module 1 – Basic Analytics

This module contains **6 scenario-based quiz questions** which revolve around a common data scenario with only 1 dataset (i.e., 1 table in 1 CSV file). The expect average solve time for all questions within this module is between 10-15 minutes.

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

- Data aggregation/summarizing, such as reporting the group of data points with highest sum or average
- Computations using the entire dataset, such as computing percentages across groups of data points
- Simple comparison across multiple groups of data points, such as comparing sums/averages
- All questions 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

- Explicitly instructing candidates to perform a certain data operation
- Data merging/joins, such as working with multiple datasets/input tables
- Data tasks that do not require manipulating the entire dataset, such as locat-

ing a specific row, column, or value

- Statistical analyses and inference
- Complex data transformations

Module 2 – Databases & SQL

This module contains **1 coding question** focused on databases and SQL, with an average solve time of 15-25 minutes.

Expected Knowledge

- Basic SQL operations, functions, and queries, 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

- Common SQL queries in popular SQL-based databases (i.e., MySQL, PostgreSQL, Microsoft SQL)

Should Exclude

- Complex queries which require deeper knowledge of SQL operations and functions not discussed above
- Tasks which require handling NoSQL or non-tabular data (i.e., JSON, XML, graph).

Module 3 – Advanced Analytics

This module contains **8 scenario-based quiz questions** which revolve around a common data scenario with only multiple datasets (i.e., 3-5 tables across CSV files). The expect average solve time for all questions in this module is between 25-30 minutes.

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

- Merging data from multiple tables together, then aggregating or summarizing the combined data to extract insights
- Computations using the multiple data tables
- Comparing across multiple groups of data points using appropriate statistical analyses and inferences
- All questions should revolve around a data scenario with multiple dataset/ input tables of medium-large size (500-2000 rows) that are easily accessible via most data analytics tools

Should Exclude

- Explicitly instructing candidates to perform specific data operations, such as computing average for a specific group/class within the dataset
- Data tasks that do not require manipulating multiple datasets, such as locating a specific row, column, or value
- Complex data transformations
- Complex statistical analyses or machine learning algorithms

Framework Example Content

Below are example questions for each module of the framework[1]. Similar questions are consistently being developed in accordance with framework specifications and monitored on an ongoing basis to minimize the impact of potential leaks that could result in cheating or plagiarism, ensure the reliability and validity of evaluations, and provide relevant and fair candidate experiences through changing industry standards.

Module 1 – Basic Analytics

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:
[Candidates are asked to download the dataset (data2.csv)]

Data explanation (in the first 2 rows):

date	time_slot	total_user_time_spent_in_mins	total_ads_watched_in_mins	ads_clicked	user_time_spent_versionB_in_mins	ads_watched_versionB_in_mins	ads_clicked_versionB

[1] Example questions are for reference only, and examples may not match the exact number of questions outlined for each module in the framework.

2020-06-01	00:00-05:59	409671	26195	22	299059	24164	1

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 Questions:

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?
2. Which time range has the smallest percentage of watched ads over all the ads watched?

Module 2 – Databases & SQL

Scenario

The elf factories have already finished producing all Christmas presents, but before Santa 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 Santa's sleigh, 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 Christmas 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.

Example

For the following tables **gifts:**

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

and **packages:**

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

the output should be:

package_type	number
big	2
medium	1
small	1
special	1

Module 3 – Data Collection

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 Questions

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.
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 `α = 0.05`.