Technical test in Data Science

Hello candidate. Congratulations on reaching this stage! We look forward to seeing more of your ability to analyze, transform, and process data. Any questions that arise during the process, please feel free to contact recrutamento@datasprints.com. Shall we?

Goals

The purpose of this test is to assess your proficiency in the **minimum requirements** for our job opportunity:

- Basic SQL programming;
- Basic Python programming;
- Cloud Computing experience;
- Linux experience;
- Statistics, Machine Learning, Optimization and Data Mining fluency;

Also, **differentiation skills**:

- Ability in analyzing large volumes of data;
- Knowledge in building data pipelines;

Therefore, focus on showing that you **meet our criteria** by answering the minimal requirements using a mix between *Python and SQL*. Leave to cover the differentials in the points raised in bonus items.

Deadline

You will have **7(seven) days** to submit this test **from the date this email was sent**. We consider that you have at least 1(one) weekend of effort to complete it. If you want more time, please contact us to set a deadline that best fits us all. :)

About the data

We provide you with data from the **NYC Taxi Trips** database. There are three different sets of files:

Dataset	Descrição
Trips (1, 2, 3 e 4)	Taxi Trips data in New York City.
Vendor Lookup	Taxi Service Company Data
Payment Lookup	Map between prefixes and actual payment types

Minimum requirements

From the data provided, we want you to answer, preferably with graphs, the following questions:

1. What is the average distance traveled by trips with a maximum of 2 passengers;
2. Which are the 3 biggest vendors based on the total amount of money raised;
3. Make a histogram of the monthly distribution over 4 years of rides paid with cash;
4. Make a time series chart computing the number of tips each day for the last 3 months of 2012.

Bonus items

These are things that we will be glad if you do, and they will definitely set you apart from the other candidates:

- What is the average trip time on Saturdays and Sundays;
- Analyse the data to find and prove seasonality
- Create assumptions, validate against a data and prove with storyelling and graphs
- Find what the fare amount (inclusive of tolls) for a taxi ride in New York City given the pickup and dropoff locations
- Make a latitude and longitude map view of pickups and dropoffs in the year 2010;
- To be able to provision your entire environment in a public cloud, preferably AWS.

data **sprints**

How to deliver

You must deliver to us 3(three) essential packages:

- **README.md -** file containing instructions for reproducing your analysis;
- **Analysis.html** - file with the answers to each question from the minimum requirements section and with the additional reviews from the bonus section. When possible, use storytelling techniques to make your thinking clear to us;
- Source code: The codes and queries you used to build your analysis. (extra points for deliveries in repositories such as **Github or Gitlab**).