Data Science Assignment

Brief

At Spare, we're interested in understanding how you think about situations that you might face while working here. This is why we have developed a Data Science assignment - to give you the opportunity to show off how you approach a problem, and to give you a sense of some of the work we tackle every day. You should take **no more than 2 hours** to complete this assignment.

As a Data Scientist at Spare, you will be responsible for uncovering insights from our data, and communicating them clearly and imaginatively to a variety of internal and external stakeholders. We would like to see you flex these skills as much as possible in this assignment!

The assignment is not about the technical aspects of Spare's platform. Although it touches on some basic transportation concepts, your response should not require outside research.

Scenario

Spare's software powers several on-demand transportation services in Minnesota, USA. Minnesota Viking Transit (MVT) – a fictitious agency – has been using Spare's software for several months and has noticed that riders are cancelling their trips relatively frequently. MVT has a hunch that certain groups of riders are cancelling their rides more often than others.

MVT would like to better understand what groups of riders are cancelling more frequently, why this might be the case, and what they can do to reduce cancellations among those groups.

As a Data Scientist at Spare, it's your job to help answer these questions using the data you have available to you!

Please briefly present your methods, findings and recommendations to Minnesota Viking Transit in a short report (1–2 pages). Be prepared to explain your methods and findings at a subsequent Spare interview.

Hints and tips about the data

You are provided with **two datasets.**

1. **The first dataset,** *'SampleDataset_Jan2020_Aug2020'*, is a sample of all trips taken on Minnesota Viking Transit between the months of January and August 2020. Each row of the dataset represents a single trip. The columns contain information you could use to stratify your data. The **password** to access this dataset is **7quibble137!**

    - **requestedPickupTs**: the requested pickup time of the trip, in UNIX timestamp format
    - **requestedDropoffTs:** the requested dropoff time of the trip, in UNIX timestamp format
    - **requestedPickupTsTimestamp:** the requested pickup time of the trip, in datetime format
    - **requestedDropoffTsTimestamp:** requested dropoff time of the trip, in datetime format
    - **requestedPickupLatitude:** the latitude of the requested pickup location of the trip
    - **requestedPickupLongitude:** the longitude of the requested pickup location of the trip
    - **requestedDropoffLatitude:** the latitude of the requested dropoff location of the trip
    - **requestedDropoffLongitude:** the longitude of the requested dropoff location of the trip
    - **status:** an indicator of whether the trip was completed or cancelled
    - **travelDistance:** the total theoretical travel distance of the requested trip
    - **travelDuration:** the total theoretical travel time of the requested trip
    - **endpoint:** the way in which the trip was booked:
        - 'nextAvailable' means the rider booked a trip on MVT's app to leave as soon as possible
        - 'scheduled' means the rider booked a trip on MVT's app at least 2 hours in advance of their trip (up to 2 days in advance)
        - 'flagDown' means the rider physically flagged down an MVT vehicle they saw driving down the street.
2. **The second dataset,** '*SampleDataset_Census_Minnesota'* contains interesting census data from Minnesota that could help boost your analysis. Each row contains the data for a single Census Block Group (a standard census division in the US). We have given you the latitude and longitude of the centroid for each Census Block Group, which you could use to geospatially join with the trip data. Each Census Block Group also has sociodemographic information per column (with units separated from the column title by an underscore, '_'). Think about how you could use this census data to define different 'rider groups' for the analysis you present to MVT.

In the interests of time, you can assume these datasets are as 'clean' as they need to be for this analysis.

Next Steps

Please send us your **short written report in PDF format.** It should include a brief overview of your methodology, as well as key findings and your chosen data visualizations. Feel free to include code as an appendix if you wish.
Please send this to alina@sparelabs.com via Google Drive.
Once we have reviewed your assignment, we will reach out to discuss the next steps.

Good luck!