CBDS Data Scientist
Capstone Project

Overview:

Congratulations! You've just finished the CBDS Blitz Training and Practice Leads are eager to put you to work. Luckily, we have an opportunity that your training has prepared you for. Please note – this project is not a collaborative effort. You must complete this project on your own.

Your client, The Mayor of New York City, needs a better understanding of Citi Bike ridership. He wants an Operating Report for the Year of 2017 on his desk by the end of the week. Based on previous engagements we know the mayor is a big fan of visualizing data in charts.

Luckily, Citi Bike publishes quarterly trip data available for you to download and analyze. The data includes:

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name
- End Station Name
- Station ID
- Station Lat/Long
- Bike ID
- User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member)
- Gender (Zero=unknown; 1=male; 2=female)
- Year of Birth

Specifically, the Mayor wants to see a variety of data visualizations to understand

1) Top 5 stations with the most starts (showing # of starts)
2) Trip duration by user type
3) Most popular trips based on start station and stop station)

4) Rider performance by Gender and Age based on avg trip distance (station to station), median speed (distance traveled / trip duration)
5) What is the busiest bike in NYC in 2017? How many times was it used? How many minutes was it in use?

Additionally, the Mayor has an idea that he wants to pitch to Citi Bike and needs your help proving its feasibility.

He would like Citi Bike to add a new feature to their kiosks: “Enter a destination and we’ll tell you how long the trip will take”.
We need you to build a model that can predict how long a trip will take given a starting point and destination. You will need to get creative about the factors that will predict travel time. For example, weather and traffic patterns may have an impact on Citi Bike travel time. There is certainly data out there – you just have to find it.

Grading Criteria

1) Ability to obtain the Citi Bike data for 2017.
2) Ability to create a 2017 Citi Bike Operating Report showing visualizations the following:
    a. Top 5 stations with the most starts (showing # of starts)
    b. Trip duration by user type
    c. Most popular trips based on start station and stop station)

    d. Rider performance by Gender and Age based on avg trip distance (station to station), median speed (trip duration / distance traveled)
    e. Busiest bike in NYC in 2017
    f. Want to really impress the Mayor? Try building something like this: https://secretnyc.co/video-pulse-new-york-city-visualized-using-citi-bike-gps-data/
3) Ability to transform / prepare data for consumption (merge, remove duplicates, etc.)
4) Create a model that can predict travel time based on a starting point and a destination. Include variables that may have an impact on travel time.
5) Presentation. Create a deck with your results. Treat this as an opportunity to present your results to the client. Slides should have concise titles, charts should be labeled and legible, and you should clearly answer the client's questions.

Submission Details

Once completed please submit your Operating Report and code used to analyze data (in Git or a notebook) to the BringIt Challenge.
You need to make sure you code is easily understood. For each component of the dashboard you should write specific procedures for how someone should read your code. You should provide detailed explanations in the README file and your code should be commented and clean.

Further submission details:

1. Read Me: this is the walkthrough of what you did. This should be a Word document with numbered steps describing your workflow. You should answer all the questions in the project description here, including a description of your model. Make sure to note any assumptions you made in the process.
2. Presentation: this is the version of your results you would present to a client. include your visualizations, answers the client's questions, and make decisions about what the client needs to know.

3. Your code: This should be where you actually did you work, in whatever tool you used (Jupyter Notebook, RStudio code, SPSS stream, etc). This should be well commented so a reader could understand what you did without needing to run it.

More tutorials to try:

Analyze precipitation data

Analyze Facebook Data Using IBM Watson and Watson Studio

Watson Assistant Workspace Analysis with User Logs