**CBDS Data Scientist**

**Capstone Project**

**Overview:**

Congratulations! You’ve just finished the CBDS DevUp 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 Chicago, needs a better understanding of Divvy Bike ridership and the restaurants operating in Illinois. 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, Divvy Bike publishes quarterly trip data available for you to download and analyze. The data includes:

* trip\_id: ID attached to each trip taken
* start\_time: day and time trip started, in CST
* stop\_time: day and time trip ended, in CST
* bikeid: ID attached to each bike
* tripduration: time of trip in seconds
* from\_station\_name: name of station where trip originated
* to\_station\_name: name of station where trip terminated
* from\_station\_id: ID of station where trip originated
* to\_station\_id: ID of station where trip terminated
* usertype: "Customer" is a rider who purchased a 24-Hour Pass; "Subscriber" is a rider who purchased an Annual Membership
* gender: gender of rider
* birthyear: birth year of rider
* latitude and longitude for stations

Also Yelp provides information related to the performance of restaurants for you to download and analyze. You can use the Yelp open data rather than the dataset challenge data. The dataset has multiple files but for the purposes of our analysis restrict your usage to the following files:

* Check in
* Reviews
* Business

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 Chicago in 2017? How many times was it used? How many minutes was it in use?
6. Find the top 10 and bottom 10 restaurants in Illinois having most and least check ins respectively.
7. For the top 10 and bottom 10 restaurants calculated in step 6, calculate the average star rating and average sentiment score of the reviews
8. What are the top 10 Cuisine types (Mexican, American, Thai, etc) based on the number of restaurants and number of check ins
9. What are the most popular keywords or adjectives that reviewers use for the above list of cuisines (calculated in step 8)
10. Are there temporal trends (daily, weekly, holidays) associated with business check ins?

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

He would like Divvy 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 Divvy Bike travel time. There is certainly data out there – you just should find it.

Also, the Mayor has a couple of hypothesis which he needs you to verify with the given data. “Do you find a correlation between the star rating and length of reviews?”

The Mayor also wants us to identify the topics from the reviews of restaurants and explore if these topics are insightful and whether they can be used as inputs to some other classification or predictive work in the future. To do this learn a set of topics by applying topic modeling algorithms such as LDA on textual reviews. Choose the number of topics in a data-driven fashion such as by using a figure that plots perplexity versus number of topics.

**Grading Criteria**

1) Ability to obtain the Divvy Bike data and Yelp data for 2017.

2) Ability to create a 2017 Operating Report showing the visualizations that were asked above.

Note : 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) Approve or Disapprove the hypotheses that the Mayor has using statistics

6) Ability to identify the main topics from the reviews of business from the reviews of restaurants using NLP.

7) 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 follow these steps to submit:

1. Create a box folder that contains:

• Operating Report

• ReadMe

• Your code

Note: The graders will need to access your box folder, in order to have the correct permissions, please follow these steps to ensure you are providing the shared link for the folder: [Capstone Box Folder Creation](https://ibm.box.com/s/s5iq3xf6l3rz5xluy5qvsqd47n99g0t9)

2. Once your box folder is created and the documents have been uploaded. Please fill out the submission form <https://ibm.box.com/s/he91ja8ck4k78yp5sd1wywmrz37b0r7h> and include your shared box link.

3. To submit, email your Data Science Submission document ONLY to cptnaedu@us.ibm.com. You should not be sending any additional files.

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 about the specific files you’ll be including in Box:

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 did your 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](https://dataplatform.cloud.ibm.com/exchange/public/entry/view/5a7051906b8fe9cc1ba126b53ed945ee)

[Analyze Facebook Data Using IBM Watson and Watson Studio](https://dataplatform.cloud.ibm.com/exchange/public/entry/view/93f7d6ac0d2d073f0d54c204f3a3e9d7)

[Watson Assistant Workspace Analysis with User Logs](https://dataplatform.cloud.ibm.com/exchange/public/entry/view/a757b0929af5c0cf5196d03d3d4064fc)