# PyBer Challenge

## Overview of Project

###Original Project

In this project, I worked as a data analyst for PyBer, a Python based ride-sharing app company. My assignment was to perform an exploratory analysis on data in some very large CSV files. To aid in the process, I created several types of visualizations to tell a compelling story about the data. I wrote Python scripts using Panda’s libraries, the Jupyter Notebook, and Matplotlib, Python’s two-dimensional plotting library, to create a variety of charts that showcase the relationship between the type of city and the number of drivers and riders as well as the percentage of total fares, riders, and drivers by type of city. The analysis and visualizations that I created helped PyBer improve access to ride-sharing services and determine affordability for underserved neighborhoods. Omar, my supervisor at PyBer, provided a series of practices for me to learn the Matplotlib library. This included creating line charts, bar charts, scatter plots, bubble charts, pie charts, and box-and-whisker plots, and making them visually compelling and informative by adding titles, axes labels, legends, and custom colors. Omar also introduced me to SciPy, a statistical Python package, and NumPy, a fundamental package for scientific computing in Python. I used Pandas, SciPy, and NumPy to perform summary statistics. </p>

### Revision of the Project

V. Isualize, the CEO of PyBer, gave Omar and me a brand-new assignment. Using my Python skills and knowledge of Pandas, I created a summary DataFrame of the ride-sharing data by city type. Then, using Pandas and Matplotlib, I created a multiple-line graph that shows the total weekly fares for each city type. This allows our audience to absorb this information quickly and detect patterns, trends, correlations, and outliers in the data. Our visualizations and findings can be shared across teams to help project planning and drive decisions. Finally, I prepared this written report that summarizes how the data differs by city type and how those differences can be used by decision-makers at PyBer. </p>

## Challenges

Throughput working on Omar’s practice series on the Matplotlib library, I found the Python code did not compile. Like the PyCitySchools project, sometimes finding the code errors was simple, such as just running the entire code from the beginning rather than just the cell. However, sometimes it was much more challenging to find the error. In some of these situations, I referred to another analyst who had previously worked on a similar assignment, WANGHEN21, “Pyber\_challenge.ipynb.” In others I used PyCharm to solve the coding error.</p>

## Results:

### PyBer Ride Sharing Data Summary Table

The following table provides a summary of the PyBer ride sharing data.



![PyBer\_Summary\_Table.png](https://github.com/Robertfnicholson/PyBer\_Analysis/blob/7672dd56a1e51ac607962f52e97839fc78759b49/Resources/analysis/PyBer\_Summary\_Table.png).

Urban areas had 13.0 times more rides than under-served rural areas and 2.6 times more rides than suburban areas. Urban areas also had 111.0 times more potential drivers than rural areas and 7.0 times more potential drivers than suburban areas. For total fares urban areas had 9.2 times more total fares than rural areas and 2.1 times more total fares than suburban areas. The order of magnitude between urban areas and suburban and rural areas decreased compared to total rides as more potential drivers in urban areas decreased the fares charged and / or fewer drivers in rural and suburban areas resulted in higher fares in those city types. This is demonstrated in both the average fare per ride and the average fare per driver. Rural areas had the highest average fare per ride and urban areas had the lowest with suburban areas falling between rural and urban. Rural drivers received the highest fares, 12.0 times higher than urban drivers and 3.6 times higher than suburban areas. It is unclear the impact of distance traveled or time of each ride on ride sharing fares as this was not part of the data set. </p>

### PyBer Ride Sharing Total Fare By City Type Multiple Line Graph

The following multiple line graph provides total fare by city type:

![Pyber\_fare\_summary.png](<https://github.com/Robertfnicholson/PyBer_Analysis/blob/9c11d20f3af177504a65b3323e5f91e88375ff63/Resources/analysis/PyBer_fare_summary.png>).

The graph provides a visualization of the total fares to show the relationship between city type and total fares. The graph provides a different perspective than the summary table. Unlike the summary table, the graph provides fare total by city type **by day** over the full timeframe of the data series. This shows the relationship between the city types and total fares daily. </p>

## Key Reports

In addition to the above table and graph, I generated the following key charts for PyBer’s CEO as part of the project:

* PyBer Ride- Sharing Data (2019) bubble chart

![Fig1.png](https://github.com/Robertfnicholson/PyBer\_Analysis/blob/9c11d20f3af177504a65b3323e5f91e88375ff63/Resources/analysis/Fig1.png)

* Ride Count Data (2019) box and whisker chart

![Fig2.png](<https://github.com/Robertfnicholson/PyBer_Analysis/blob/9c11d20f3af177504a65b3323e5f91e88375ff63/Resources/analysis/Fig2.png>)

* Ride Fare Data 2019 box and whisker chart

![Fig3.png](<https://github.com/Robertfnicholson/PyBer_Analysis/blob/9c11d20f3af177504a65b3323e5f91e88375ff63/Resources/analysis/Fig3.png>)

* % of Total Fares by City Type pie chart

![Fig5.png](https://github.com/Robertfnicholson/PyBer\_Analysis/blob/9c11d20f3af177504a65b3323e5f91e88375ff63/Resources/analysis/Fig5.png)

## Summary

Recommend the following to address the disparities among the city types:

* Get distance traveled or total time of each ride in order to examine the impact on driver count and fares.
* Publish the findings upon the CEO’s review and feedback to assist our stakeholders including riders and drivers. The findings may prompt drivers to cover rural and suburban areas given the higher fare rates.
* Complete a follow up analysis that includes distance traveled or total time of each ride and publish these results.</p>