# **New York City Taxi Fare Prediction**

# **Dan Siegel spring 2019 GitHub Portfolio URL**

# Which Domain?

What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you’re doing with these data.

1. <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8> Logistic regression step by step
2. <https://developers.google.com/maps/documentation/> the google maps API documentation
3. <https://towardsdatascience.com/bayesian-linear-regression-in-python-using-machine-learning-to-predict-student-grades-part-2-b72059a8ac7e> Bayesian Regression to predict the taxi fares
4. <http://www.greenteapress.com/thinkbayes/html/index.html> Think Bayes is a good overview on Bayes theorem, and has a good amount of code available
5. <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1> Overview of K-means clustering
6. <http://benalexkeen.com/k-means-clustering-in-python/> real code examples of Kmeans clustering
7. <https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1> random forest classifier guide
8. <https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm> good in depth dive into random forest
9. <https://medium.com/datadriveninvestor/a-simple-guide-to-creating-predictive-models-in-python-part-1-8e3ddc3d7008> A good overview on in depth data cleanup and prep
10. <https://www.pyimagesearch.com/2016/08/22/an-intro-to-linear-classification-with-python/> overview of linear classification

# Which Data?

What is the dataset you’ll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.

<https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data>

# Research Questions? Benefits? Why analyze these data?

How are you proposing to analyze this dataset? This is about your approach. Here, you’ll be proposing your research questions as well as justifications for why you’d offer these data in this way.

The research question I have posed for myself, is to see if additional data will impact the analysis. I am going to complete a very similar analysis to those of the other participants of the competition. The key thing here is that I am seeking to work with the Google Maps API, and in conjunction with the time of day, seeing if the distance and trip duration make an impact on the predicted outcome. I will be using Google API to get the estimated time and distance of the trips and adding them in to the models.

# What Method?

What methods will you be using? What will those methods provide in terms of analysis? How is this useful?

I will be using the Google API, and testing other users methods and a number of predictive methods.

# Potential Issues?

What challenges do you anticipate having? What could cause this project to go off schedule?

This is a for fun challenge, I have no idea of the accuracy of the data set is true. I do not know if there are real latitude/longitude combinations. I also have the free tier of Google API, so that will affect the number of calls I am able to perform.

# Concluding Remarks

This is a good example of going outside your provided data set and generating new data. Is this data useful? Does it add value? Does the time taken to side track into this extra data befit the goal or will it not increase the accuracy at all.