Assignment 4

# **Overview**

For this assignment you'll do a realistic exploration of data up to and including creating predictive regression model(s).

# **Data**

The data is the included datasets:

* **Linear Regression**: "NYC\_Real\_Estate.csv", which is a subset of some real data on the housing market in NYC. You'll use this dataset to create a linear regression model.
* **Logistic Regression:** "credit\_risk\_dataset.csv", which is a set of simulated credit data. You'll use this dataset to create a logistic regression classification model.

# **Part 1 - Data Preparation: Both Datasets**

See the two previews above

Note: There's nothing that you explicitly need to produce here, it is getting your data ready for the next step.

The first part of your process should be to clean and prepare the data (each set). It is in its original form, so there's some stuff in there that isn't perfect. Some things to consider are:

Are the variables in the data relevant? Some may not be. Use both general knowledge and analysis to make a judgement.

Are there missing values? If so, what does that mean? Especially with the T/F columns in the data.

What data format is the data? Categorical/numerical?

Are there outliers? Should they stay?

There's no expectation for you to do any extreme amount of research or analysis to use the data, the cleanup required should be reasonably simple, but required. What to keep/remove/change is up to you. There is not one fixed answer. As you make choices, note a justification. This preparation and clean up kind of blends into the EDA for each.

# **Part 2 - Linear Regression - Real Estate Data: Sales Price (35% of grade)**

You'll do a linear regression to predict the **SALES PRICE.** As part of this you'll need to do some EDA (exploratory data analysis) which will involve walking through exploring and visualizing the data to make sense of what is there, what is useful, and what is not.

# **Part 3 - Logistic Regression - Credit Data: Loan Status (35% of grade)**

You'll do a logistic regression to predict the **LOAN STATUS**. As part of this you'll need to do some EDA (exploratory data analysis) which will involve walking through exploring and visualizing the data to make sense of what is there, what is useful, and what is not.

# **Part 4 - Are the Models Good (15% of grade)**

Please briefly note if the models you've created are good. Justify your answer.

**Overall Format (15% of grade)**

Please make your workbook readable. Someone should be able to read through from end to end and get a solid understanding of what happened and what the findings were. The goal is to basically say, "here's a walkthrough of how I did this regression, and here are the results."

You Formatting is important in making things clear. This shouldn't end up being insanely complex, if it does, you've likely overthought it in some way.

Data Dictionary and Documentation<

For the real estate data, there are some notes in the real\_estate\_notes.txt file in this repository that explains some of the data values. The PDF is the 'glossary of terms" that the notes file refers to.

For the credit data, there are some notes in the credit\_notes.txt file. This one is really simple in terms of what the data is and means

I don't think that you need much background information to make sense of this data, it has been simplified a bit. If something doesn't make sense, please ask.

Tools, Technologies, and Techniques

You can use whatever tools/packages you'd like to do the regression and the EDA. Thinkstats, thinkplot, seaborn, scipy, scikitlearn, statsmodels, etc are all fair game, as are anything else. You'll just need to import them.