**Data Science Program Final Project**

**Zillow title here**

**Executive Summary**

At the end of the Data Science program, students are required to complete a final project of their choice. They are given six weeks to work on the project.

This document is dedicated to my final project. It will explain the purpose and scope for the project.

**Business Objectives**

To showcase the skills that I have acquired through Entity’s Data Science program, I’ll be using R, Python, Tableau and other programs to wrangle, analyze, and visualize the “Housing Prices” dataset made available by UCI Machine Learning on Kaggle, along with census data from the Census Bureau.

At the end of the project, I will explain my work in layman’s terms, and present findings to students, faculty, staff, and potential employers, along with other interested parties via Zoom.

**Background**

As one who renovates houses in my spare time, I chose the dataset because I am interested in real estate, machine learning and the ability to predict housing prices in different markets, along with delving into possible reasons behind Zillow’s recent failed attempt last year to add home-flipping to their portfolio. The real estate listing giant was confident in its ability to leverage artificial intelligence to estimate the value of homes but just 8 months later shut down that division after losing nearly $500 million. The challenge they faced was being able to accurately forecast future prices of homes in a market that rapidly became volatile due to Covid, a shortage of licensed contractors, rising material costs, to name a few. Machine learning is good for predicting the past, or for predicting stock prices a second or two in advance, but predicting housing prices 3 to 6 months ahead of time may be a task that “byte” off more than it can process.

I hope to gain some insight from this investigation in order to better understand the real-world factors that can affect machine learning alogrithms.

**Scope**

I will be using the software taught in the program to complete the project. I plan to focus on the most sought-after programming languages: Python, R, Tableau, Trello and use PowerPoint and Prezi or genial.ly, or a combination of these to present my data story in an engaging way. I may use SQL to join additional databases, just to demonstrate proficiency.

**Functional requirements**

Data Wrangling: The downloaded dataset will need to have subsets for some data wrangling. Columns, null values and unusable columns will be removed for these subsets. The datatypes for each column will be converted to a usable format for the needed analysis in each case using R and Python.

Data Analysis: I will explain what each column means, and how the values are measured. I have put together a list of questions to ask, and what they might gather from the dataset that can reduced as I gain insight from preliminary analysis. I will narrow the scope of information in the dataset if it is unnecessary. Then, I will identify the proper functions to create models, predictions, etc.

Data Visualization: Once I have a comprehensive understanding of and insight gathered from the dataset, I will work on visualizing the findings. I may use Tableau or other graphing programs, and compile the visuals and texts in a Power Point, Prezi or genial.ly presentation. I am using Trello to track the progress of my work.

Presentation: Working with school leaders, I will present my findings via Zoom at noon on Monday February (TBD). The presentation will be around 20 minutes.

**Personnel requirements**

I am the sole developer on this project. I have weekly meetings with my instructor and my coding mentor to problem-solve or to check in on work as it progresses. I need to spend roughly 15-20 hours per week on this project so that it is the best it can be.

**Delivery schedule**

Week 1: Import dataset into preferred software to begin data wrangling. Any unnecessary columns should be removed. Educate myself on the particulars of machine learning models for real estate. Update Trello. There are 79 independent variables: are they all necessary to the analysis?

Week 2: Study the dataset and ask questions. What are some possible correlations? Is the data normally distributed? What are some predictive models we can make from it? Visualize the data to discover any interesting findings. Update Trello. Contact public information office of the Census Bureau with relevant questions and attend Census training. Conduct correlation tests, ANCOVA, ANOVA, Exploratory Factor Analysis.

Week 3: Modeling/Optimization (Combined Stepwise - Forward and Backward Selection) and Machine Learning (Random Forest). Update Trello.

Week 4: Review and validate findings from the previous week, and draw insights/conclusions. Set up GitHub. Begin compiling findings into a presentation.

Week 5: Completion of compiling findings into a presentation. Go over it with my instructor, mentor and a friend to ensure that the presentation is clear and logical. Work on the style and layout of the presentation so it is engaging, easily understood and flows.

Week 6: Make final touches to the Power Point presentation. Finalize Trello, update GitHub and practice presenting.

**Other requirements**

All programs used should be free of charge.

**Assumptions**

The software programs and platforms should be available, up-to-date, and not broken.

**Limitations**

If something should come up for me during this six-week period, the project may be delayed. If the instructor or mentor have scheduled or unscheduled time-off, the project may be delayed as well. If I experience a roadblock in the work, it may push back the completion date. I work 40 hours a week, so I’ll have to make sure to schedule time on the weekends and Monday and Tuesday evenings after my instructor and mentor meetings to work on the project.

**Risks**

The risks that may arise are such like natural disasters, power outages, family emergencies or broken software/hardware. I am excited about the subject material and the opportunity to test my skills on a new set of data, and motivated to complete the program so there should be no completion issues. The instructor and mentor are phenomenal so there is no concern of no help from them. The support and enthusiasm I’ve received from them has been so motivating. The risk of this project being incomplete is infinitesimal. I will be successful in completing this project!