## Course Project Instructions

One of the main objectives of this course is to help you gain hands-on experience in communicating insightful and impactful findings to stakeholders. In this project you will use the tools and techniques you learned throughout this course to both find insights and assess the quality of a data set that you feel passionate about.

After going through some guided steps, you will have a data set in great shape, ready for hypothesis testing, and ready for supervised or unsupervised machine learning. You will also produce a report that helps you focus on highlighting your analytical skills and thought process.

Project

### *Optional: find your own data set*

As a suggested first step, spend some time finding a data set that you are really passionate about. This can be a data set similar to the data you have available at work or data you have always wanted to analyze. For some people this will be sports data sets, while some other folks prefer to focus on data from a datathon or data for good.

### *Optional: participate in a discussion board*

As an optional step, go into a discussion board and brainstorm with your peers great data sets to analyze. If you prefer to skip this step, feel free to use the Ames housing data set or the Churn phone data set that we used throughout the course.

### **Required**

Once you have selected a data set, you will produce the deliverables listed below and submit. Treat this exercise as an opportunity to produce analysis that are ready to highlight your analytical skills for a senior audience, for example, the Chief Data Officer, or the Head of Analytics at your company.

Sections required in your report:

1. Brief description of the data set and a summary of its attributes
2. Initial plan for data exploration
3. Actions taken for data cleaning and feature engineering
4. Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner
5. Formulating at least 3 hypothesis about this data
6. Conducting a formal significance test for one of the hypotheses and discuss the results
7. Suggestions for next steps in analyzing this data
8. A paragraph that summarizes the quality of this data set and a request for additional data if needed

***You will be required to submit your report in .pdf format.***

*Make sure to include mainly insights and findings on your report. There is no need to include code, unless you want to.*

## **FAQs**

Here are five frequently asked questions about the assignment and review process. Read these before starting your assignment.

Do I have to come up with my own data set?

* You are highly encouraged to find a data set you feel really passionate about. This will help you showcase analytical work that truly matches your skills. But if you prefer, you can use some of the data sets from this course.

Is it OK to choose the same data set as someone else?

* Yes, more than one person can analyze the same data set. Most likely your insights will be different from your peers and you will still be able to showcase your own talent as a unique solution.

Should I save the data after I have performed data transformations, treatments, and cleaning?

* It is highly recommended that you save your clean data set. You might use it later in some exercises for supervised and unsupervised machine learning.

Is this an individual assignment?

* You can ask for help or assistance on technical issues and general direction of your analysis, but the interpretation of the analytical output and the writing of the report should be your own.

1. Brief description of the data set and a summary of its attributes

The house prices: advanced regression techniques from kaggle

Housing prices

Msssubclass: identifies the type of dwelling involved in the sale. From size, number of floors, number of rooms, etc.

1. Initial plan for data exploration

Looking through the data and identify variables and looking for null values, sizes, and number of non-numeric values, identifying outliers. Using: .describe(), .info(), .head(), dtypes()

1. Actions taken for data cleaning and feature engineering

Checking for skewness. The predictions generated by the final model will be log-transformed, so I'll convert these predictions back their original form later on. I will start with making a histogram with plt.hist(). Then I will transform train.salesprice with np.log() and calculate the skewness and plot the data again. Thenn, I will check for correlations between the features and the target with .corr()followed by using .unique() to get the unique values. Then to further look at the relationship between ground living area and sales price with scatter plots as well as with garage area.

1. Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner

There is a positive correlation between sales price and overall quality, ground living area, garage cars, and garage area. Using a pivot table as well as a scatter plot, I visualized the correlation between sales price and overall quality.

Using feature engineering and plotting a bar plot of the correlation between street features and sales price. This shows that partial has a significantly higher median sales price.

1. Formulating at least 3 hypotheses about this data

- my hypotheses are the larger the house the higher the sales price

- the better the overall quality the higher the sales price

- the larger the lot area ground living area the higher the sales price

1. Conducting a formal significance test for one of the hypotheses and discuss the results

Testing hypothesis of the influence of lot area on sales price

H0: the earlier built date of the house result in the lower house price

H1: the earlier built date of the house does not result in the lower house price

Alpha to be set to 0.05, which is a 95% confidence level.

I'll calculate a Pearson correlation coefficient and the p-value for testing correlation from scipy.

The result of the p-value suggest we should reject the null hypothesis.

1. Suggestions for next steps in analyzing this data

Building a linear model, that will separate the features and the target variable for modeling.

1. A paragraph that summarizes the quality of this data set and a request for additional data if needed

Because this is a data set from Kaggle intended for beginners, I found the data to be well organized and complete.

kaggle kernels pull code/ryanjt/house-prices

## Course Project Self-Review

Self-Evaluate your course project submission.

1. The summary of the data helps clearly understand the size of the data set and its variables, including possible target variables.

* No points awarded if there is no summary or it is hard to put together what variables are available. (0 pts)
* One point if there is a basic summary, like a data dictionary. (1 pt)
* One extra point if the summary of the data is presented by types of attributes. (2 pts)

2. Does the report include a plan for data exploration that seems sound, makes sense, and highlights some of the vision for the analysis?

* This report is missing a planning section for the data analysis (0 pts)
* Yes. This plan includes a detailed subtask section or a good vision of what is possible to do with this data set. (1 pt)
* This plan exceeds expectations. In addition to plan out subtasks and vision for this analysis, it also anticipates possible snags that might be incorporated into preliminary hypothesis of the data. (2 pts)

3. The report thoroughly discusses the results of the Exploratory Data Analysis? Is it informative, actionable, and insightful?

* No. Some output is presented but there are no clear findings, insights, or takeaways (0 pts)
* Yes (1 pt)

4. Does the report clearly describe any actions taken for data cleaning and feature engineering?

* No. There is no section that highlights any steps taken to ensure the quality of the data. (0 pts)
* Yes, some actions were taken to address the quality of the data, like missing value treatment, encoding, or visualizations. But there is still room for improvement (.5 pts)
* Yes, some actions were actions taken to address the quality of the data, like missing value treatment, encoding, or visualizations. But there was a key part missing, and it will be pointed out. (1 pt)
* Yes, several actions were actions taken to address the quality of the data, like missing value treatment, encoding, or visualizations. There is a clear sense of before/after treatment and evidence of the data being in great shape. This section exceeds expectations. (2 pts)

5. The report contains a section with clear key findings and insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner.

* No. The report does not include this section (0 pts)
* Yes. The report includes an Insights section but that seems rushed or not too insightful. (1 pt)
* Yes, the report contains a section that thoroughly discusses clearly-written key findings and insights derived from the Exploratory Data Analysis. EDA was performed and presented above expectations. (2 pts)

6. The report contains a section that formulates at least 3 hypothesis relevant to this data set?

* No. The report is missing this section or it presents hypothesis that seem irrelevant to this data. (0 pts)
* Yes. The report includes a section that formulates 3 good hypothesis relevant to this data. (1 pt)
* Yes. The report includes a section that formulates 3 or more great hypothesis relevant to this data. Testing one or more of these hypothesis for significance will be really insightful. (2 pts)

7. Does the report discuss thoroughly the results of a significance test at least one of the hypothesis in an insightful manner?

* No. The report does not include any formal test for significance of a given hypothesis. (0 pts)
* Yes. The report thoroughly discusses a significance test for at least one good hypothesis. The results are presented in an insightful and actionable manner even if there are slight misinterpretations or room for feedback. (1 pt)
* Yes. The report thoroughly discusses a significance test for at least one great hypothesis. The results or their presentation come across as truly insightful and above expectations, even if there are slight misinterpretations or room for feedback. (2 pts)

8. Does the report include conclusions and next steps?

* No, this section is missing. (0 pts)
* Yes, this section is there with good insights. (1 pt)
* Yes, this section is there with good insights and helpful, actionable next steps. (2 pts)

#### If you scored less than 8 points, please review the lessons taught in this course again.