**Individual Project by Anjesh Sahani**

**Topic and Dataset Collection**

**Contextual Research and Data Description**

While topic context and data collection were high-level exercises requiring only a few sentences - here I want you to expand and really explain the variables and the measures used for those variables. I need to see a demonstration of understanding your dataset and what you might be able to find in it – but also what might be missing and need more expansion later.

Introduction – background and context by exploring the data in the following ways:

1. **What is the domain from which the dataset is drawn?**

* I have taken this dataset from **Kaggle competitions.**
* **Link: -** [House Prices - Advanced Regression Techniques | Kaggle](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview/description)

1. **Why did you choose this dataset?**

* The reason I have chosen this dataset because it required Advance **Regression** **Techniques** and feature engineering work before even doing any analysis.

For example:

first, I have to clean the data by looking each column, splitting the column inside column, changing categorical variable into numeric, dropping columns that doesn’t explain dataset before analysis etc.

1. **What are the variables in the dataset? (what are they named, what do they measure, and how do they measure it)**

* There are 81 different variables in this dataset, each column has specific meaning for this data. here I am mentioning some variable name. If you want in details about all the 81 columns, then please [**Tap-here**](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data)**.**
* **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.
* **MSSubClass**: The building class
* **MSZoning**: The general zoning classification
* **LotFrontage**: Linear feet of street connected to property
* **LotArea**: Lot size in square feet
* **Street**: Type of road access
* **Alley**: Type of alley access
* **LotShape**: General shape of property
* **LandContour**: Flatness of the property
* **Utilities**: Type of utilities available
* **LotConfig**: Lot configuration
* **LandSlope**: Slope of property
* **Neighborhood**: Physical locations within Ames city limits
* **Condition1**: Proximity to main road or railroad
* **Condition2**: Proximity to main road or railroad (if a second is present)
* **BldgType**: Type of dwelling
* **HouseStyle**: Style of dwelling
* **OverallQual**: Overall material and finish quality
* **OverallCond**: Overall condition rating
* **YearBuilt**: Original construction date
* **YearRemodAdd**: Remodel date
* **RoofStyle**: Type of roof
* **RoofMatl**: Roof material

1. **What are you thinking (or hoping) you will find?**

* My main **goal** is to **predict** the **sales price for each house** based on multiple test model and ultimately decide which model will be best fit to explain the dataset.
* **Example**: - Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence
* Where I will get to know about each variable and relationship between them.

1. **What is missing?**

* There are multiple columns that have **NULL, Missing** values, which need further data cleaning and manipulation work.
* Example: I have to deal with NULL value, first I have to set a maximum limit if any column has more than MAX limit then that column will be dropped, or I have to fill with mean or mode value according to nature of variable.

1. **Why is this topic meaningful to you?**

* **Because** this topic will boost my Analysis knowledge, reason this topic not just only require data visualization work but also require data cleaning, machine learning model, feature engineering, training and testing the **Model**.

I want to know what your topic means to you, but I also want you to show that you understand the components of your dataset and have a grasp of what your story will be.

**Grading – 25 Points**

1. 2 points for domain and how you found the data
2. 5 points for why this dataset over a different one
3. 5 points for in depth description of the variables
4. 5 points for what you are trying to discover or prove
5. 3 points for what gaps your data might have and what you might need to add
6. 5 points for explaining why this topic is meaningful or important to you (why you chose it)

**Exploratory Analysis**

Submit your data visualizations (Excel/Tableau/PowerBI/R) as a file. If you utilized dashboards I want the dashboard file, if you used to excel, I want the excel file, if you used R I want the R file. I can utilize these files to see how you created the visualizations you used to complete your exploratory analysis.

Upload a copy of your exploration write up. Preliminary data exploration and understanding your data is a critical step in any research process. This write-up should demonstrate your understanding and show me how you used the dataset that you had previously submitted to begin understanding your topic.

Your write-up should include how you went about exploration and your initial findings:

1. Data Preparation - check for zeros, null values, error values, extreme outliers, duplicates, or any other inconsistencies.
2. Visualizations - based on the variables you select create and explain your visualizations
   1. I want you to understand and look at every single variable in your data set and visualize it in some way, although you do not need to submit every visual you created
   2. Explain every variable you visualized and why
   3. Explain every variable you did not visualize and why
3. Observations and Insights - point out what you initially noticed, why you think it is interesting, and what you want to explore more
   1. What trends did you notice?
   2. What subsets stand out as interesting?
   3. Here, further exploration means either continuing to scope out or addressing the problem/opportunity

**Grading – 70 points**

1. 10 points for submitting visualization files
2. 5 points for describing how you prepared/checked your data before analysis
3. 10 points for describing your visualizations (what variables are included, why you chose the visualization method you did)
4. 10 points for describing the variables you did not visualize and why
5. 35 points for describing what initial observations stood out to you and what you plan to further explore about.... **and why**

**Explanatory Analysis**

Submit your data visualizations (Tableau/PowerBI/R) as a file. If you utilized dashboards I want the dashboard file, if you used R I want the R file. I can utilize these files to see how you created the visualizations you used to complete your exploratory analysis.

Note: I am looking to ensure you can create aesthetic visualizations - I will give you notes on aesthetics and quality of the visualizations to help for your final materials. I will take off points for poorly constructed visualizations. Also note, you **may not use excel** for this portion.

Explanatory analysis is where you take all the data, all the research, and all the knowledge you have and you apply it to drive your story and justify your position. Additionally, transformation is critical - you **must** demonstrate that you have altered/aggregated/advanced your data in some way through your own skill and ability.

1. What problem statements or questions are you answering/solving for?
2. Transforming the dataset - explain the new variables you have created
   * Define them (what do they mean)
   * How did you make them? (aggregation/filtering/outlier-removal/etc.)
   * Why did you make them? (what value did they provide to your analysis)
3. For each graphic/visualization explain the following:
   * What is the purpose of the chart (how does it help your story/what is it telling the audience)?
   * Why did you choose each individual graphic type to represent your data?
   * How did you manipulate or what did you do to the data to create the variable/s you are choosing to represent?

**Grading – 50 points**

1. 10 points to establish what problems or opportunities you discovered
2. 10 points to describe the new variables you have created (including how)
3. 10 points for why those variables add value to your analysis (how you used them in explanatory analysis)
4. 10 points to explain the charts and graphics you will use for your explanatory analysis
5. 10 points for how your graphics either solve or do not solve for the problem/opportunity (i.e. what did you learn - what was the outcome?)

**Differences between assignments and final paper**

You will have similar sections – however what I am looking for in each will be very different.

**Intro/Context:**

**Assignment:** I want to know why you chose the topic (why is it meaningful to you), how you found your sources and data, what you are learning along the way. You can/should present what you think you will find, but there is no points off for not being totally sure, being wrong, or changing your hypothesis later. This is about what you are finding and how you found it.

**Final:** I do **not** want a deep dive in why this topic is interesting to you. This is a more formal submission. You need to explain the context so a reader could reasonably understand the topic background to ease their understanding of the rest of the paper (i.e., it’s written with the rest of the paper in mind). You need to present what your argument will be and how you are going to explain it. This is about telling the reader what to expect from the paper.

**Data Preparation:**

**Assignment:** Describe every variable every time I ask. Demonstrate you understand the variables you are working with and what they mean. You **must** identify and explain all the ones in the datasets you leverage. Understanding data sources and what a data table contains is an absolutely critical skill to any analytical work. If you cant interpret the data you are using effectively then you could taint your entire project and ruin the legitimacy of your findings.

**Final:** You only need to describe the variables that you use in your visualizations and that you ran analysis on. You do not need to include other variables that you do not end up using in your final submission. HOWEVER – bonus points to anyone who addresses possible criticisms or issues with their analysis and explains why they chose **not** to include certain variables. Understanding and addressing objections early is a great way to add legitimacy to your point. I wont specifically take away from your final papers if you do not do this. But I will 100% add points you might have missed earlier if you do this effectively.

**Exploratory Analysis:**

**Assignment:** The book uses an example of oyster diving for pearls. Exploratory analysis is searching 100 oysters for two pearls. In your assignment I want to know about the 100 oysters you are looking through. What did you do to with the data, what analysis did you run, what charts did you use, what did you find told you something, and what did you find that told you nothing. I want perspective into what your process is/was. I want details, examples, and how you did the work.

**Final:** If you tell me about every single step you ran and every single thing you found along the way I will hammer your final grade. Here you should be highlighting some initial observations and what indicated the problem/opportunity to you. Back to the oyster analogy – here I don’t want how you found all 100 oysters I want to know “I searched 100 oysters and noticed that blue oysters by a heat vent seemed to turn up more pearls than everywhere else”.

**Explanatory Analysis:**

**Assignment:** Similar to exploratory I want to know what you did, why you did it, what made it more valuable, what worked, what you thought would work but did not… I want to know why you chose and created new variables as well as why you chose to neglect or remove others. I want why you chose certain graphics, how they were better than others. I want insight to your thinking.

**Final:** I don’t want to know all they why you did it, I just want to know why it is important. Going back to the oysters – “Blue oysters by the heat vent generated more pearls. The added warmth of the water decreased the number of predators in the area and was conducive to this breed of oyster growing undisturbed which increased pearl productivity by x%”.

**Conclusion**:

There is no assignment portion for this. This will only be in the final. But it should build off your explanatory analysis – “The increased productivity of the blue oyster species could lead to developing more commercially viable oyster farming. Based on our findings these 5-6 locations would be ideal because x, y, z allowing for increased pearl output expanding the availability of gem quality pearls on the market”.

**Final Paper**

**Introduction**

Explain why you chose the topic you did and provide some additional research and context outside of your dataset that helps establish your topic and provide a "level-set" for your audience (in this case the reader - me, but this will be good practice for your final presentations

Final Paper

**Introduction and Context**

**Exploratory Analysis**

Data Preparation

What were your initial observations?  
Are there any trends or seasonality in your dataset?  
Data preparation – do you have to do any cleaning up of the raw data?  
Share preliminary insights as you explore the dataset.

1. Data quality checks
2. Detect interesting subsets to form hypotheses for hidden information.
3. Look for errors, does the data make logical sense, duplicate rows, spelling errors, and/or any missing variables?

**Explanatory Analysis**

**Conclusion**