# Relationships to Copper

Curtis L. Leavitt

Western Governors University

# Table of Contents

## A. Project Highlights

Research Question:

How do the prices of stock market indices, gold, silver, natural gas, and crude oil influence copper prices and can we build a model to predict future copper prices based on these relationships?

Project Scope:

The scope of this project will include a model that will take a new dataset and accurately predict copper prices. The scope also includes details of what variables correlate with copper price fluctuation.

Solution Overview tools:

Python was utilized for all manipulation of data, from downloading, preprocessing, building our model, and prediction. Jupyter Notebooks was used as the environment for code development, execution, and visualizing data. Various Python libraries were used including Seaborn, Numpy, Pandas, Sci-kit Learn, Scipy, XGBoost and Matplotlib.

Solution Overview Methodologies:

I used the waterfall methodology for this project, following a structured process that included requirements gathering, system design, implementation, verification, and deployment. The requirements phase involved defining key metrics, data sources, and project scope. In system design, I selected charts, the predictive model, and planned data preprocessing and feature engineering. During implementation, I built and trained the model, tuning it through hyperparameter adjustments and evaluating performance. In the verification stage, I tested the model with new data to ensure accuracy, and since the model wasn't intended for real-world deployment, no ongoing maintenance was needed.

## B. Project Execution

Project Plan:

The goal for this project was to create a model that accurately predicts copper price and highlights correlation of variables. The plan formulated in task 2 was executed accordingly.

- Goal 1: Create a model that accurately predicts copper prices.
  - o Objective 1.1: Determine what variables correlate with copper price.
    - ▪ Deliverable 1.1.1: The deliverable for this objective is charts & graphs indicating which variables highly(>.85) correlate.
  - o Objective 1.2: Provide a model that predicts copper price.
    - ▪ Deliverable 1.2.1: Model with $R^2$ score > 0.90.
    - ▪ Deliverable 1.2.2: Model with MSE < 0.01.

Project Planning Methodology:

I implemented the waterfall methodology for the duration of this project. This includes requirements gathering, system design, implementation, verification(testing), deployment, and maintenance.

**Requirements Gathering:** In this step I outlined the project scope based on stakeholders' expectations, defined key metrics for correlations and the predictive model. I also documented resources including data sources, target accuracy, and reporting needs.

**System Design:** In this step I chose the charts and graphs used to describe correlating variables. I also chose which predictive model meets the criteria in the previous step. I determined where and how the neccesary data was gathered. The data preprocessing (handling missing data, normalization, etc.) and feature engineering was determined here as well.

**Implementation:** In this step the model was built and implemented. The data was collected and preprocessed. The model was trained and validated using a training set then  hyperparameters were used for improvement. The model performance was then evaluated using key metrics.

**Verification(Testing):** In this step I verified the model accurately predicts copper price. I used new data to evaluate the general performance and ensure it meets metric requirements.

**Maintenance:** This stage was not applied, as the model was not to be used in any real-world applications requiring ongoing support. The methodology outlined in task 2 was utilized unchanged from the original plan.

Project Timeline & Milestones:

| Milestone or deliverable | Projected Start Date | Projected End date | Actual End date |
|---|---|---|---|
| Develop scope & requirements | 9/23/2024 | 9/24/2024 | 9/24/2024 |
| Choose data source | 9/24/2024 | 9/24/2024 | 9/24/2024 |
| Choose visual methods & model | 9/24/2024 | 9/26/2024 | 9/24/2024 |
| Determine data preprocessing/feature engineering | 9/24/2024 | 9/25/2024 | 9/24/2024 |
| Collect data | 9/26/2024 | 9/26/2024 | 9/24/2024 |

| | | | |
|---|---|---|---|
| Preprocess data & train model | 9/26/2024 | 9/26/2024 | 9/25/2024 |
| Visualize variable correlation | 9/27/2024 | 9/27/2024 | 9/26/2024 |
| Tune model & predict | 9/28/2024 | 10/01/2024 | 9/26/2024 |
| Evaluate model using metrics | 10/1/2024 | 10/1/2024 | 9/26/2024 |
| Test model | 10/2/2024 | 10/2/2024 | 9/27/2024 |

The timeline and milestones were followed closely with the exception of being ahead of schedule for some of the tasks. These tasks were overestimated for the amount of required time to complete them.

## C. Data Collection Process

Data Selection and Collection:

The data was downloaded via a .csv file from https://www.kaggle.com/datasets/saketk511/2019-2024-us-stock-market-data . The collection plan was utilized unchanged from task 2 outline.

Data Collection Obstacles:

The dataset acquired for this project was of good quality and seemed to be accurate and represented the variables well. There were some errors and required manipulation to utilize the dataset for model training and visualization. The data was gathered using proper methods. The completeness appeares to have some missing data for various dates to include all weekends, holidays, and any dates not recorded. There were null values, special characters where not necessary, and missing values. These issues could have caused a skew in the results of the prediction model and were taken into consideration when building the model. Some of these issues were not outlined in task 2 as they were found upon further investigation.

Data Governance Issues:

For Data Goverance, the data was managed in accordance with Kaggle's guidelines, with proper attribution to the original data sources. For Privacy, no PII or human data was present, privacy concerns are not relevant. For security, the dataset is publically available. Standard data precaution was taken, such as storage using a password protected system and proper backup procedures to prevent data loss or unauthorized access. For Ethical, Legal, and Regulatory Compliance, the dataset was used in strict accordance with Kaggle's terms of service and licensing restrictions respected. The Data Governance plan used was unchanged from the outlined plan in task 2.

### C.1 Advantages and Limitations of Data Set

The advantage of this dataset was simplicity due to it already being gathered and formatted into a csv file for use. The limitations of this dataset were that it required some preprocessing for it to be utilized properly with my model. There were 39 available variables and contained errors within some of the fields which required cleaning prior to use.

## D. Data Extraction and Preparation

The extraction for this dataset was simple. I downloaded it from the Kaggle website and saved it to my local drive. From there it was read into a dataframe using Pandas library in Jupyter Notebooks with Python. The issues with this data that required preprocessing included, unneccesary columns, columns with incorrect data (special characters), incorrect datatypes, null values, zeros, and negative numbers (should have been positive numbers). After creating a dataframe there were 2 unneccesary columns that were dropped due to being unusable and containing special characters. The datatypes incorrectly labeled as objects were converted to floats to properly represent the numbers in the fields. The “Date” column was updated from an object to a datetime object for better utilization. The null values were updated by replacing them with the mean value of the column. This was done to preserve the integrity of the data and remove the need to delete the rows with missing data.  The fields that had zeros and negative numbers were updated by replacing them with the mean of the column. This was done to ensure the integrity of the data since these numbers returned an error when recorded. Feature engineering was also utilized to add the day of the week. This was added by extracting it from the datetime object and making 5 columns, one for each day of the week (Mon-Fri).

## E. Data Analysis Process

### E.1 Data Analysis Methods

To test the hypothesis that copper prices are greatly influenced by gold, silver, crude oil, natural gas, and stock indices, the Kendall Tau correlation was used as the analytical method. A low positive Kendall Tau P-value and high positive T coeffecient will indicate a strong influence, supporting the hypothesis, while lower values may suggest weaker relationships. Kendall Tau is an appropriate non-parametric statistical test for measuring the strength and direction of association between variables, particularly when assessing ordinal or ranked relationships. This method will calculate the correlation between copper prices and each of the other commodities and indices. Providing a measure of how well changes in one variable predict changes in copper prices. This approach ensured alignment with the hypothesis by quantifying the directional associations between copper prices and its potential influencers.

### E.2 Advantages and Limitations of Tools and Techniques

Python was utilized for all manipulation of data, from downloading, preprocessing, building the model, and prediction. Jupyter Notebooks was used as the environment for code development, execution, and visualizing data. Various Python libraries were used including Seaborn, Pandas, Sci-kit Learn, Numpy, Scipy, XGBoost, Joblib, and Matplotlib. The advantages to using Python is the versatility from processing to machine learning and having an extensive library ecosystem . The disadvantages in using Python is that it is slower compared to lower-level languages like C++ and requires higher memory usage. Jupyter Notebooks advantages are being able to execute code cell by cell and allowing debugging to be easier. A disadvantage would be confusion of execution order when cells are run out of order and cause unexpected outcomes.

### E.3 Application of Analytical Methods

I applied the Kendall Tau coefficient to assess the strength of correlation between copper prices and various predictor variables like gold, silver, crude oil, and stock indices. The Kendall Tau method requires that the variables be ordinal or continuous, and that no assumption of a linear relationship is necessary, making it robust against non-normal distributions and outliers. To verify the suitability of this method, I ensured all variables were continuous and did not expect a strictly linear relationship. Additionally, I calculated the Tau coefficient and the corresponding P-values to assess statistical significance, ensuring that the results met the required criteria for meaningful correlation.

## F Data Analysis Results

### F.1 Statistical Significance

My null hypothesis is that there is no correlation between copper prices and gold, silver, crude oil, natural gas, and stock indices. The statistical test used was the Kendall Tau correlation. The outputs of this test will be a Kendall Tau coefficient(T) and a P-value. The *alpha* value used will be 0.05(5%). If the Tau coefficient > 0.7 and P-value < *alpha* then the null hypothesis is rejected. The conclusion drawn from this test was that the variable 'Ethereum_Price' showed to have a Tau coefficient $\geq$ 0.7 and a p-value < *alpha*. There is sufficient evidence to reject the null hypothesis and support the claim that there is a relationship between this variable and copper price.

For the copper price prediction, the model used was a supervised regression model. XGBoost was used to accurately predict copper prices and was considered successful with a mean squared error < 0.01 and $R^2$ score > 0.90. I was able to build a model using hyperparameter tuning with the best score of a mean squared error = 0.005025 and $R^2$ score of 0.989651.

**F.2 Practical Significance**

The practical significance is that a company that consumes copper could use these insights to optimize copper inventory and pricing stratgeies and take advantage of anticipated increases or decreases of copper price. If a consumer can know that there will be a large increase in the copper price they could mitigate risk associated with this volatility and make more informed decisions about future purchases and allocations. I found that the first model I built didn't have as much practical application due to having 39 variables. This would have made it difficult to use in the future due to needing to record a high volume of information to get an accurate prediction. I recreated the model using only 6 variables which in turn increased the mean squared error to 0.0067, making the prediction less accurate. Using this model with 6 variables would allow more pratical application for a consumer to predict the cooper price and take advantage of this savings by shifting the ordering of material.

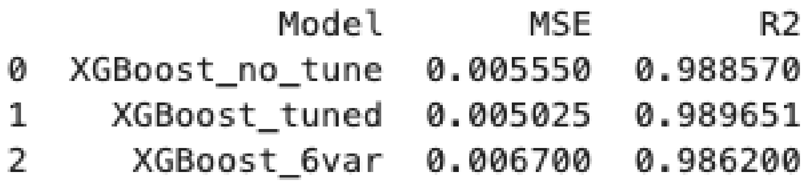
### F.3 Overall Success

With all of the criteria met, I view this project as a success. I was able to receive a P-value and T coeffecient once running a Kendall Tau statistical test which disproved my null hypothesis with a variable having a T coefficient ≥ 0.7 and a p-value < 0.5. I was able to provide charts that visualized the correlation of variables. I also was able to build a model that accurately, within parameters, predicted the copper price. The model met the criteria of having a mean squared error ≤ 0.01 and an $R^2$ score ≥ 0.90.

## G. Conclusion

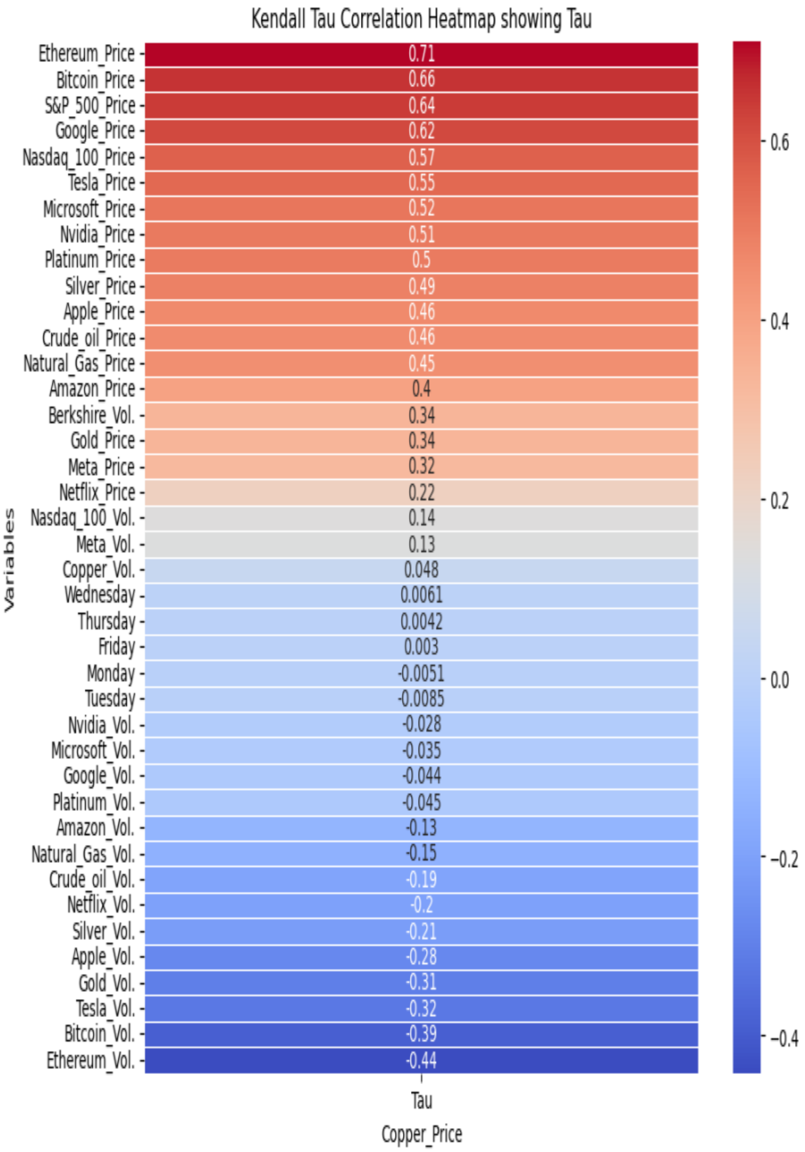
### G.1 Summary of Conclusions

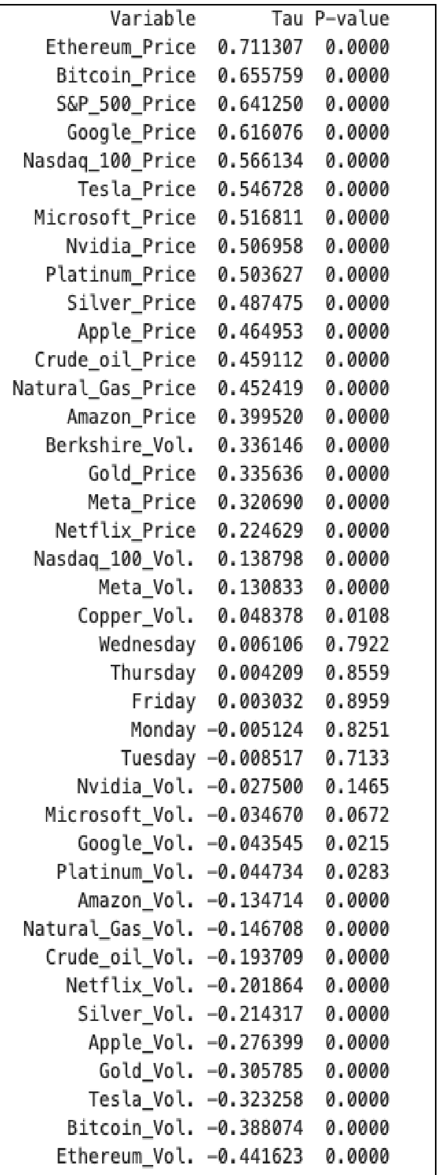
This project was meant to create a way to predict copper prices for consumers to forecast and better allocate time and resources to saving money. The image below shows the results of the models built saved as a dataframe for display. This shows that objective 1.2 deliverable 1.2.1 and 1.2.2 was achieved by all versions of the model.



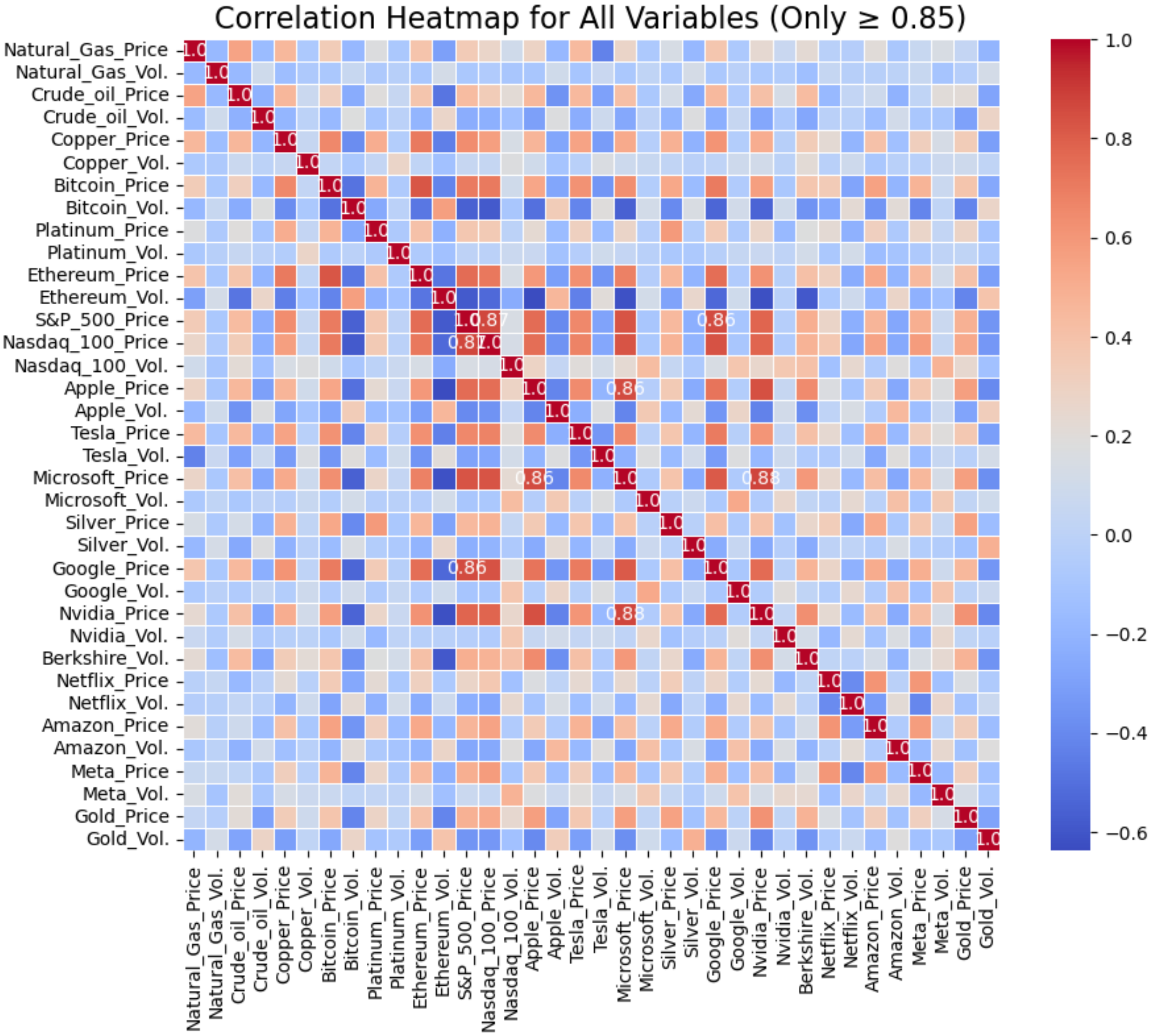
```
             Model       MSE        R2
0  XGBoost_no_tune  0.005550  0.988570
1    XGBoost_tuned  0.005025  0.989651
2     XGBoost_6var  0.006700  0.986200
```

The heatmap below visualizes the results of the Kendall Tau correlation which had a goal of a Tau coefficient ≥ 0.7 and a p-value < 0.05. Ethereum_Price was the only variable that met this criteria. The image also shows all the resulting Tau coefficients and P-values.





```
         Variable       Tau  P-value
   Ethereum_Price  0.711307   0.0000
    Bitcoin_Price  0.655759   0.0000
    S&P_500_Price  0.641250   0.0000
     Google_Price  0.616076   0.0000
 Nasdaq_100_Price  0.566134   0.0000
      Tesla_Price  0.546728   0.0000
  Microsoft_Price  0.516811   0.0000
     Nvidia_Price  0.506958   0.0000
   Platinum_Price  0.503627   0.0000
     Silver_Price  0.487475   0.0000
      Apple_Price  0.464953   0.0000
  Crude_oil_Price  0.459112   0.0000
Natural_Gas_Price  0.452419   0.0000
     Amazon_Price  0.399520   0.0000
   Berkshire_Vol.  0.336146   0.0000
       Gold_Price  0.335636   0.0000
       Meta_Price  0.320690   0.0000
    Netflix_Price  0.224629   0.0000
  Nasdaq_100_Vol.  0.138798   0.0000
        Meta_Vol.  0.130833   0.0000
      Copper_Vol.  0.048378   0.0108
        Wednesday  0.006106   0.7922
         Thursday  0.004209   0.8559
           Friday  0.003032   0.8959
           Monday -0.005124   0.8251
          Tuesday -0.008517   0.7133
      Nvidia_Vol. -0.027500   0.1465
   Microsoft_Vol. -0.034670   0.0672
      Google_Vol. -0.043545   0.0215
    Platinum_Vol. -0.044734   0.0283
      Amazon_Vol. -0.134714   0.0000
 Natural_Gas_Vol. -0.146708   0.0000
   Crude_oil_Vol. -0.193709   0.0000
     Netflix_Vol. -0.201864   0.0000
      Silver_Vol. -0.214317   0.0000
       Apple_Vol. -0.276399   0.0000
        Gold_Vol. -0.305785   0.0000
       Tesla_Vol. -0.323258   0.0000
     Bitcoin_Vol. -0.388074   0.0000
    Ethereum_Vol. -0.441623   0.0000
```



The heatmap above is visualizing the variables that show a correlation of > 0.85. There were no variables that showed a correlation greater than 0.85 with the copper price. This shows deliverable 1.1.1 of objective1.1 . This project was a success due to the fact that I was able to receive a P-value from a Kendall Tau test disproving my null hypothesis, showed correlation of variables, and provided a model that predicted copper price. The metrics for success were met however objective 1.1 deliverable 1.1.1 was not met due to no variables highly correlating (> 0.85).

### G.2 Effective Storytelling

The visualizations I provided were all built using Python in a Jupyter Notebook (ipynb file). The heatmaps were built inside the Jupyter Notebook using Python libraries Seaborn and Matplotlib. The heatmaps show a strong representation of the strength of correlations between two variables. The dataframe image represents the evaluation results of all the models built. The scores were saved together in a dataframe for better side by side comparison. The print image of the Kendall Tau results showed how all the variables scored in ascending order to compare how each variable faired in the test.

### G.3 Recommended Courses of Action

The first course of action would be to gather additional data to provide a longer timeframe than the dataset entails. The dataset currently has roughly 3 years of data, if there was a longer data history I believe the model would be more accurate in it’s prediction. The second course of action would be to utilize multiple datasets to locate variables that better correlated with copper price in order to improve the prediction. Variables could be isolated from each dataset and combined to have highly correlated variables and a more complete dataset. A third course of action would be to gather a longer timeframe of data and utilize a Time Series model to possibly better capture the changes in price due to seasons, holidays, and yearly flucuations.

## H Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=87891ad0-c8b0-4bad-878f-b1fa014edd14

# References

No sources were cited.

# Appendix A

## Evidence of Completion

https://github.com/cleav84/Relationships-to-Copper.git

1. Link to Github repository containing all files for this project:
    a. Jupyter Notebook containing Python code
    b. Kaggle Dataset
    c. Test Dataset
    d. Model saved as pkl file