**GWORLDSOFT SOLUTIONS LIMITED**

DATA SCIENCE AND ANALYTICS PROJECT

***Project 1: Stock Market Trend Analysis Using Yahoo Finance.***

***Introduction:***

Stock analysis refers to the method that an investor or trader uses to evaluate and investigate a particular trading instrument, investment sector, or the stock market as a whole. Stock analysis is also called equity analysis or market analysis. Investors or traders make buying or selling decisions based on stock analysis information.

Stock market analysis is crucial for investors to make informed decisions. This project involves analyzing historical stock prices of Apple(AAPL), Bitcoin(BTC-USD), Google(GOOGL), Facebook(META) and Tesla(TSLA), identifying their trends and evaluating risk through statistical and visual methods. Python programming language was used to fetch, clean, and analyze stock data using the moving average indicator.

***Analysis***

The technical analysis method involves examining data generated through market activities, such as volume and prices. Analysts

following such a type of stock analysis use technical indicators and tools like charts and oscillators to identify patterns that can indicate future price trends or direction.

Technical analysts examine the historical trading data of a security and estimate the future move of the security. It is frequently used for forex and commodities. The technical analysis is based on the following assumptions:

* **Price follows a trend.** It implies that once a trend is established, future prices tend to follow the direction of the trend. Such an assumption is the basis of many strategies for technical trading.
* **History is likely to be repeated.** History repeats itself, mainly concerning price movement. [Market psychology](https://www.dailyfx.com/education/understanding-the-stock-market/stock-market-psychology.html) causes price movements to repeat. Technical analysis involves using chart patterns to analyze the movements in the market and study trends. Charts that have been used for over 100 years are still relevant since price movement patterns are often repetitive.

***Diagrams:***

The following charts were generated to visualize trend and correlation of the selected stock:

1. Figure 1: Stock Prices the stocks over the last 5 years.
2. Figure 2: Stock prices of the AAPL stock of the last 5 years.
3. Figure 3: Stock prices of Tesla over the last 5 years.
4. Figure 4: Stock prices of Google over the last 5 years.
5. Figure 5: Price trend of Facebook over the last 5 year.
6. Figure 6: Line chart showing the first

***Observations***

* 1. From the charts, we observe that BTC-USD is more volatile than the rest of the selected stocks. That is why the trend of the rest of the stocks is not visible in figure 1.
  2. The 50 day rolling average of AAPL stock indicates price fall in the first quarter of 2020 and a spike in price towards the end of 2020. 200 day MA of the AAPL stock shows its minimum price around 2023 and its maximum price in early 2024.
  3. In comparing 50 day and 200 day moving average of GOOGL stock, it was observed that there was a fall in the price of GOOGL stock around 2020. This is noticed in most stock prices as caused by covid-19 pandemic. However, 200 day moving average of GOOGL stock seem not affected by the global pandemic as its minimum prices appears around 2023.
  4. The Chart of the 50 day MA of TSLA showed a current downtrend in the year 2025. 200 day moving average showed downtrend in the present year but not as steep as in the 50 day rolling average

***Conclusion***

In summary, stock trends can be observed and analyzed using moving average indicator. A single stock can exhibit different trends as thought the trend charts are of another stock. It is critical for traders as well as stock analyst to know the kind of indicators to use when initializing a trade. Moving average is one of the easiest and efficient indicators to use.

***Project 4: Customer Segmentation Using Mall Data***

***Introduction:***

This project applies clustering algorithms (K-Means) to categorize mall customers based on spending behavior. K-mean clustering is an unsupervised machine learning technique that group data points into clusters based on similarity. A k means clustering example illustrates how this method assigns data points to the nearest centroid, refining the clusters iteratively. Understanding what is k-means clustering will enhance your grasp of [data analysis](https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/) and pattern recognition.

The goal is to group similar data points together and discover underlying patterns or structures within the data.

* Recall the first property of clusters – it states that the points within a cluster should be similar to each other. So, our aim here is to minimize the distance between the points within a cluster.
* There is an algorithm that tries to minimize the distance of the points in a cluster with their centroid – the k-means clustering technique.
* K-means is a centroid-based algorithm or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid.

Businesses uses customer segmentation to tailor marketing strategies and decision making.

***Diagrams:*** The following diagrams were generated to visualize different features of the data:

1. Figure 1: Distribution plot of Annual Income in k$
2. Figure 2: Distribution plot of Age
3. Figure 3: Distribution plot showing spending score in percent
4. Figure 4: Scatter plot of annual income vs spending score
5. Figure 5: Bar plot of CustomerID Vs Spending (1-100)
6. Figure 6: Scatter plot showing the iteration of the clusters

# ***Principal Component Analysis(PCA):***

# This project also seek to implement PCA. Having too many features in data can cause problems like overfitting (good on training data but poor on new data), slower computation, and lower accuracy. This is called the [curse of dimensionality](https://www.geeksforgeeks.org/videos/curse-of-dimensionality-in-machine-learning/), where more features exponentially increase the data needed for reliable results.

The increased feature combinations makes sampling harder. In high-dimensional data, tasks like **clustering or classification more complex and slow.**

To tackle this problem, we use Feature engineering Techniques, such as **feature selection** (choosing the most important features) and **feature extraction** (creating new features from the original ones). One popular feature extraction method is [**dimensionality reduction**](https://www.geeksforgeeks.org/dimensionality-reduction/), which reduces the number of features while keeping as much important information as possible. PCA is one of the most popular dimensionality reduction method.

*Conclusion*

In this project, we discussed one of the most famous clustering algorithms – K-Means Clustering. We implemented it and looked at its step-by-step implementation. We try to assigning data points to the nearest centroid, iteratively refining clusters. K-means clustering is an unsupervised learning method that group unlabeled data into clusters based on similarity. It’s a technique to minimize distances between data points and their respective cluster centers.

PCA is a statistical technique introduced by mathematician Karl Pearson in 1901**. It works by transforming high-dimensional data into a lower-dimensional space while maximizing the variance (or spread) of the data in the new space**. This helps preserve the most important patterns and relationships in the data.

PCA helps in distilling complex data into its most informative elements, making it simpler and more efficient to analyze.

1. It identifies the directions (called **principal components)** where the data varies the most.
2. It projects the data onto these directions, reducing the number of dimensions while retaining as much information as possible.
3. The new set of uncorrelated variables (principal components) is easier to work with and can be used for tasks like regression, classification, or visualization.

***Project 5: FIFA 23 Player Performance Analysis***

***Introduction:***

Football clubs rely on data analytics for player scouting. This project analyzes player attributes, salaries, and performance metrics to identify top talents. This project focused on building predictive model that evaluates and predict players’ valuation. Ordinary least square regression algorithm was used. The sole aim is to predict the price value of the players based on some features like:

* LW Rating, LF Rating, CF Rating, RF Rating, RW Rating, GK Rating as independent features and
* Value( in Euro) as the dependent feature

***Answers to key Project Question:*** The Project question were answered as follows:

1. The player with the highest market value is Kylian Mbape. This is because he the highest potential of 95 and he is very young (23 yrs old).
2. There is a weak positive correlation between players’ age and their performance.
3. Argentina, Belgium, Poland and France are the joint top- ranked Nations in the FIFA overall ranking.
4. Players’ passing accuracy has a greater effect on their value than shooting power and speed.

**DIAGRAM**

The following diagrams were generated to communicate findings:

* Histogram of players’ potential and value
* Histogram of players’ value (in Euro)
* Boxplot of players’ age
* Bar-chart showing players and their value.

**Ordinary Least Squares regression** (**OLS**) is a common technique for estimating coefficients of [linear regression](https://www.xlstat.com/en/products-solutions/feature/linear-regression.html) equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression), often evaluated using r-squared. **Least squares** stand for the minimum **squares error (SSE)**. Maximum likelihood and generalized method of moments estimator are alternative approaches to OLS.

In practice, you can use linear regression in many fields:

* meteorology, if you need to predict temperature or rainfall based on external factors.
* biology, if you need to predict the number of remaining individuals in a species depending on the number of predators or life resources.
* economy, if you need to predict a company’s turnover based on the amount of sales.

From our analysis, the following parameters were recorded.

1. R-squared = 0.188
2. F-stats = 142.6
3. Skewness = 7.590
4. Kurtosis = 96.046

|  |  |  |
| --- | --- | --- |
| Variables | Error | P-value(t = 0.025) |
| LW Rating | -2.237e+04 | 0.195 |
| LF Rating | 1.26e + 05 | 0.000 |
| CF Rating | 1.26e + 05 | 0.000 |
| RF Rating | 1.26e + 05 | 0.000 |
| RW Rating | -2.237 + 04 | 0.195 |
| GK Rating | 2.291e + 05 | 0.000 |

***CONCLUSION:***Observe the little prediction accuracy of our model. With R-Squared of 0.188 , it means that our independent variables lacks enough power to make a more precised prediction of our dependent variables. This could be caused by multicolinearity, insufficient or poor quality of the data, inappropriate model selection, overfitting or incorrect hyperparameter tuning.

***Project 6: Google Search Trends Analysis***

***Introduction:***

Search trends reflect public interest in topics like cryptocurrency, AI, and elections. This project analyzes keyword trends over time and across regions.

Google Trends is a website that analyzes and lists the popular search results on Google search based on various regions and languages.Unofficial API for Google Trends. Allows simple interface for automating downloading of reports from Google Trends.

***DIAGRAMS***

* 1. A line chart showing the trend of each public interest.
  2. Heatmap showing the correlation analysis of the different public interest

***CONCLUSION:***

***Project 7: Airbnb Price Prediction & Market Analysis***

***Introduction:***

Airbnb hosts need pricing strategies to maximize bookings. This project predicts rental prices using regression models and also seek to identify key pricing factors.

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019. This data file includes all needed information to find out more about hosts, geographical availability and necessary metrics to make predictions and draw conclusions.

This project emphasizes on exploratory data analysis and application of random forest algorithm in solving regression problem and predict the price of travel**.**

## [**Exploratory Data Analysis - EDA**](https://www.digitalocean.com/community/tutorials/exploratory-data-analysis-python#exploratory-data-analysis-eda)

* EDA is applied to **investigate** the data and **summarize** the key insights.
* It will give you the basic understanding of your data, its **distribution,** null values and much more.
* You can either explore data using graphs or through some python **functions.**
* There will be two type of analysis. [Univariate and Bivariat](https://www.digitalocean.com/community/tutorials/autoviz-module-in-python)**e.** In the univariate, you will be analyzing a single attribute. But in the bivariate, you will be analyzing an attribute with the target attribute.
* EDA can be done using non-graphical and graphical approaches
* In the **non-graphical approach**, you will be using functions such as shape, summary, describe, isnull, info, datatypes and more.
* In the **graphical approach**, you will be using plots such as scatter, box, bar, density and correlation plots.

In this project, we will be using the two approaches as long as the desired result is achieved.

***RANDOM FOREST***

A Random Forest is a collection of decision trees that work together to make predictions. In this article, we'll explain how the Random Forest algorithm works and how to use it.

Random Forest algorithm is a powerful tree learning technique in Machine Learning to make predictions and **then we do voting of all the tress to make prediction**. They are widely used for classification and regression task.

* It is a type of classifier that uses many decision trees to make predictions.
* It takes different random parts of the dataset to train each tree and then it combines the results by averaging them. This approach helps improve the accuracy of predictions. Random Forest is based on [ensemble learning](https://www.geeksforgeeks.org/a-comprehensive-guide-to-ensemble-learning/).

***ANSWERS TO KEY QUESTION***

1. What factors (location, amenities, reviews) most influence Airbnb prices?

Amenities has the highest contribution and influence on Air prices with correlation value of 0.13

2. How does seasonality affect rental prices?

Was observed a negative correlation coefficient of -0.08 between seasonality and prices. This implies that increase in seasonality result in a decrease in the price of flight.

3. Can we build a model to recommend optimal pricing?

Sure! A regression model with prediction accuracy (r2-score) of 0.7 was built.

***DIAGRAM:*** *The following diagram were created during the analysis process:*

1. Heatmap showing the relationship between amenities, city, number\_of\_reviews and price.
2. Heatmap of seasonality and price.

***CONCLUSION:*** Random Forest learns from the training data like a real estate expert. After training it predicts house prices on the test set. We evaluate the model's performance using [Mean Squared Error and R-squared Score](https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/) which show how accurate the predictions are and used a random sample to check model prediction.

## Random Forest is very good in providing very accurate predictions even with large datasets, it can handle missing data well without compromising with accuracy. It doesn’t require normalization or standardization on dataset and it can be combined with multiple decision trees to reduce the risk of overfitting of the model.

## However random forest has Limitation of computationally cost especially with a large number of trees. Also, interpretation of the model is hard compared to simpler models like decision trees.

***Project 8: IMDB Movie Ratings & Sentiment Analysis***

***Introduction:***

Movie studios analyze ratings and reviews to gauge success. This project explores trends in movie ratings, genres, and public sentiment. The major work in the project is to clean the data, explore the data and perform sentiment analysis on the data.

***About The Dataset:*** The following are the features of the the dataset and their meaning:

* **Poster\_Link** - Link of the poster that imdb using
* **Series\_Title** = Name of the movie
* **Released\_Year** - Year at which that movie released
* **Certificate** - Certificate earned by that movie
* **Runtime** - Total runtime of the movie
* **Genre** - Genre of the movie
* **IMDB\_Rating** - Rating of the movie at IMDB site
* **Overview** - mini story/ summary
* **Meta\_score** - Score earned by the movie
* **Director** - Name of the Director
* **Star1,Star2,Star3,Star4** - Name of the Stars
* **No\_of\_votes** - Total number of votes
* **Gross** - Money earned by that movie

Sentiment Analysis in Python offers a powerful solution to this challenge. This technique, a subset of [Natural Language Processing](https://www.analyticsvidhya.com/blog/2023/09/advanced-natural-language-processing-nlp/) (NLP), involves classifying texts into sentiments such as positive, negative, or neutral. By employing various [Python libraries](https://www.analyticsvidhya.com/blog/2024/05/top-python-libraries-for-data-visualization/) and models, analysts can automate this process efficiently. Let’s delve into how to perform sentiment analysis in Python and explore some examples of its application. So, basically we share information regarding this in the article about sentiment analysis using python and how to do sentiment analysis in python.

In this project, we will focus on understanding applying Sentiment Analysis Model, including its applications and a practical Sentiment Analysis on real time data.

***Answers to Key Questions***

1. Which genres receive the highest ratings?

The genres that has the highest rating is animation, drama and war.

2. Does runtime affect movie ratings?

Movies with high runtime attracts lesser ratings. Optimum runtime attracts the highest rating and movies that are too short likewise has reduced rating

3. Can we predict box office success using metadata?

Yes!

***DIAGRAMS:*** The following diagrams were used to illustrate findings:

* 1. Barplot of positive and negative sentiment
  2. Histogram showing number of words in each review.
  3. Wordcloud showing the most frequent words in positive reviews

***CONCLUSION:***

Sentiment analysis in Python offers powerful tools and methodologies to extract insights from textual data across diverse applications. Through this project, various approaches were explored using python packages such as plotly.express, word tokenizer, wordcloud, poster-stemmer and machine learning-based models for sentiment analysis. These methods enable organizations to monitor brand perception, analyze customer feedback, and even predict market trends based on sentiment.

***Project 9: Uber Trip Analysis – Time-Series & Geospatial EDA***

***INTRODUCTION:***

Uber trip data provides insights into urban mobility patterns. This project analyzes pickup trends, peak hours, and geographical hotspots in New York City using time-series and geospatial analysis.

***About The Data:*** This dataset was created to help in exploring and predicting pickups in the area. The main dataset contained over 10 million observations of 4 variables which aggregated per hour and borough, and then joined with the rest of the datasets producing 29,101 observations across 13 variables. These variables are:

* pickup\_dt: Time period of the observations.
* borough: NYC's borough.
* pickups: Number of pickups for the period.
* spd: Wind speed in miles/hour.
* vsb: Visibility in Miles to nearest tenth.
* temp: temperature in Fahrenheit.
* dewp: Dew point in Fahrenheit.
* slp: Sea level pressure.
* pcp01: 1-hour liquid precipitation.
* pcp06: 6-hour liquid precipitation.
* pcp24: 24-hour liquid precipitation.
* sd: Snow depth in inches.
* hday: Being a holiday (Y) or not (N).

***Answers To Key Questions:***

1. What are the busiest hours/days for Uber trips in NYC? The busiest hour per day is the 19th hour with about 800 pickups.

2. Which boroughs (Manhattan, Brooklyn, etc.) have the highest demand? The borough with the highest demand is the Manhattan

borough with about 2400 pickup counts.

3. How does weather (rain, snow) affect Uber demand? Rain and snow depth has adverse effect on pickup demands. Rain affects

demands with correlation coefficient of -0.02 and snow depth affects demand with correlation coefficient of -0.008.

4. Are there seasonal variations in trip frequency? From the dataset, it was observed that demand increases steadily from 400 in

January to 500 in February and decreases slightly to about 450 in the month of march. It had been on increase from 450 in

march to more than 550 in june.

5. Can we identify popular pickup/drop-off clusters? From the dataset used, there was no noticeable point of clusters

The main **AIM** of this project is to create interactive chart using plotly in python.

Plotly is a Montreal-based AI and Analytics company. They focus on the development of Analytics tools, mainly Dash and Chart Studio. They have also released the free and open-source plotting library “Plotly” for Python, R, MatLab, and Julia.

Plotly produces interactive graphs, can be embedded on websites, and provides a wide variety of complex plotting options. The graphs and plots are robust, and a wide variety of people can use them. The visuals are of high quality and easy to read and interpret.

Plotly makes a wide variety of charts, including basic and statistical charts, maps, 3D charts, subplots, and more.

***DIAGRAMS:*** The following charts were generated to communicate findings:

1. Barchart showing hourly pickups by pickup counts
2. Bar plot of pickups by different borough
3. Correlation heatmap showing how weather factors affect demands
4. Histogram of pickups by counts
5. Line plot of pickups by months
6. Scatterplot of pickups to check clusters
7. Line plot of pickups by hours
8. Line plot showing dialy pickup trends
9. Line plot of weekly pickup trends
10. Plotly scatter Showing daily pickups
11. Plotly line of daily pickups

***CONCLUSION:***

Plotly Python library is an open-source module that is used for data visualization and supports various graphs like line charts, scatter plots, bar charts, histograms, area plots, etc. Plotly Python library produces interactive graphs, can be embedded on websites, and provides a wide variety of complex plotting options. The graphs and plots are robust, and a wide variety of people can use them. The interactivity also offers a number of advantages over static matplotlib plots, such as saving time when initially exploring your dataset.

***Project 10: World Happiness Report – GDP vs. Happiness Correlation***

***INTRODUCTION:***

The World Happiness Report ranks countries based on factors like GDP, social support, and life expectancy. This project explores how economic and social factors influence happiness scores.

***About the Dataset:***

***Answers To Key Questions:***

1. Which countries rank highest in happiness, and why? Finland is ranked the happiest nation with a ladder score of 7842

2. Is there a strong correlation between GDP per capita and happiness? There is a strong correlation of 0.79 between GDP per

capital and happiness score

3. How do social support, freedom, and corruption perceptions impact happiness? Social support and happiness score are

positively correlated with correlation value of 0.76. Freedom to make life choice also has positive correlation value of 0.61 with

happiness score. However, perception of corruption has inverse relationship with happiness score with correlation coefficient of

-0.42. This means that the better the perception of corruption, the less happier the people

4. Has global happiness improved over the years? Yes

DIAGRAMS: The following diagrams were generated to communicate findings:

1. A line plot showing happiness score by region
2. Scatter plot of happiness score by region
3. Histogram showing the distribution of happiness

CONCLUSION: It is observed in this project that the scatter plot of happiness ladder score follows a normal distribution. It histogram form a bell-shaped curve. This means the data points are clustered around the mean (average) in a symmetrical, bell-shaped curve. The majority of data values fall within a relatively small range around the mean, with fewer and fewer values occurring further away from the mean. This kind of distribution is also known as a Gaussian distribution or a bell curve.

***Project 11: Credit Card Fraud Detection – Anomaly Detection***

***INTRODUCTION:***

Fraud detection is critical for financial institutions. This project uses machine learning (Logistic Regression, Random Forest, Isolation Forest) to identify fraudulent transactions.

## ***About The Dataset***

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

The dataset contains transactions made by credit cards in September 2013 by European cardholders.  
This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount. 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

***Answers to key Questions***

1. What percentage of transactions are fraudulent?

2. Which features (amount, time, location) are most indicative of fraud?

3. How do different ML models compare in fraud detection accuracy?

4. Can we reduce false positives without missing fraud cases?

Tasks to Perform:

● Handle class imbalance (SMOTE/undersampling).

● Train and evaluate models (precision, recall, F1-score).

● Optimize thresholds for fraud alerts.

Deliverables:

● Fraud detection model (Python script).

● Confusion matrix and ROC curve visualizations.

● Report on model performance.