**Loss-function : a function which is associated with a cost in making decision eg, fastest route selection in terms of time (it is for Single Data Point) where as**

**Cost function is overall aggregate loss for entire points**

**Actual**

**P TP FP**

**Rr FN TN**

**Precission : TP / (TP+FP) eg, email not a spam but still it shows under spam**

**Recall : TP / (TP + FN) eg. when patient have covid and your model said you don't have**

**F1 : harmonic mean : when both needs to be considered eg. stock market**

**Clustering :** Silhouette Score and Dunn’s Index

**NLP - Embeding dimension** : 1536

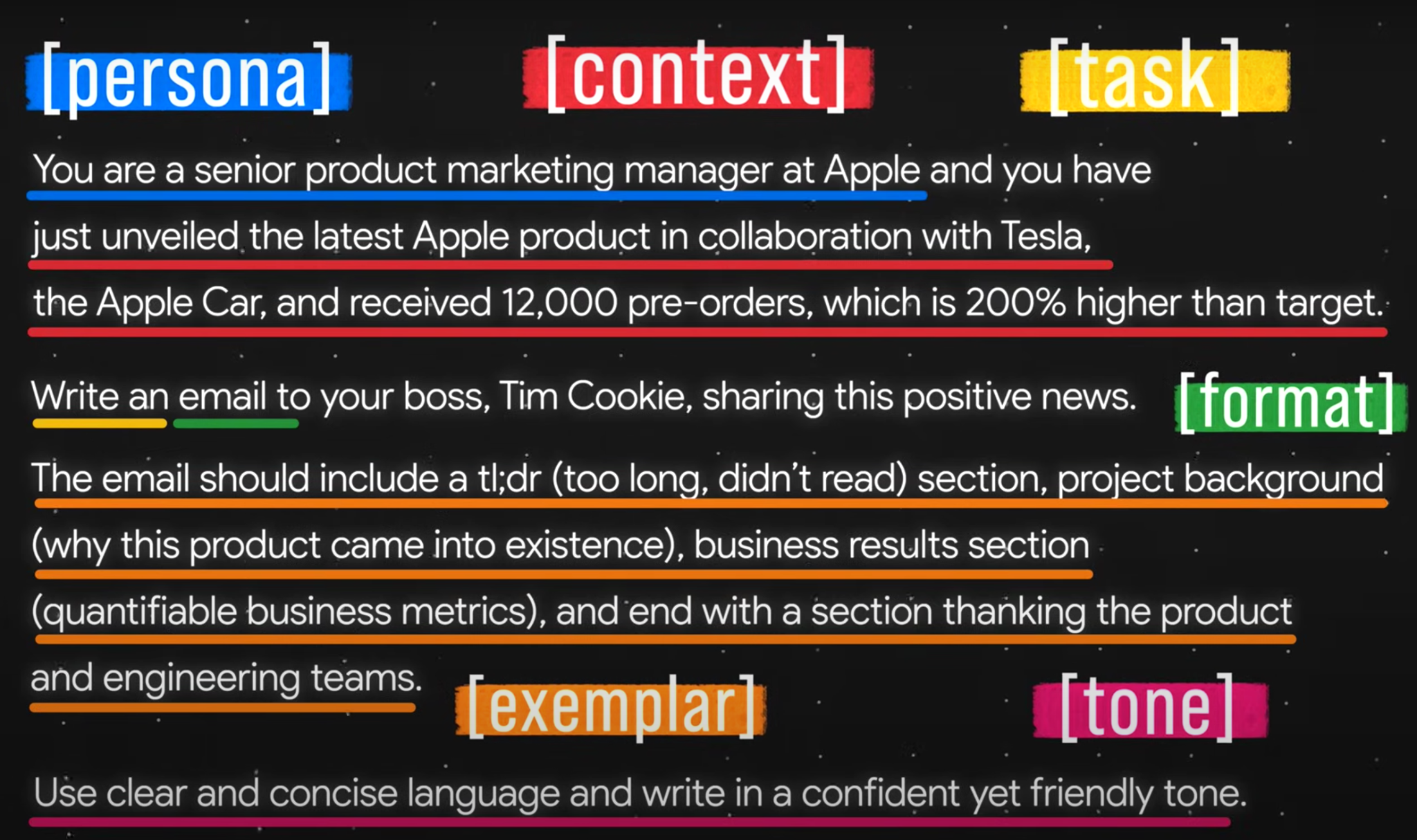
**================================================================**

**Evaluation Framework** : Big Bench, Glue Benchmark, SuperGLUE Benchmark, traditional way if sentiment analysis or classification then easy

**Prompt Types :**

1. Zero shot Prompting : No Labeled data (work dynamic task based on reasoning capability)
2. Few shot prompting : (providing a few labeled examples in the prompt) - more example more accuracy

Fine Tuning LLM: PEFT and QLoRA



BERT(Bidirectional encoder representation from transformer) :

* Transformer based architecture (Encoder+Decoder) which used unidirectional training
* Bidirectional training - Encoder only
* Goal : to understand the particular language
* Pre Training part + Finetune part
* Transfer learning
* Feedforward n/w + softmax
* BERT-BASE(784 dimentions), BERT-Large

My name is Juned Ansari, and I bring over 12+ years of comprehensive IT experience, including 7+ years in Data Science experience with a focus on machine learning, deep learning, NLP, Generative AI and advanced analytics in diverse industries.

Worked in multiple domains, including e-commerce, taxation, unemployment claims, pharma & healthcare(Abbotte, Alkem, Sun pharma, GSK-GlaxoSmithKline), logistics, travel, hotel industries, and education domain.

Comfortable with both Python and R in data science

Worked with Regression, classification, clustering, custom model building, model fine tuning, hyperparameter tuning and model selection(Grid Search CV), Generative AI, LLM’s

* I have over 9 years of team handling experience and also working as individual contributor, handled max team size of 10 individuals in the field of data science
* Roles and Responsibilities
  + Communicating with client, Lead and Manage team and development, delivery, configuration, maintenance, and support.
  + collaboration within cross-functional teams
  + I am experienced in developing end to end ML solutions using the complete CICD pipeline. Including Docker Containerization, Git, Github, Bitbucket, GithubActions.
  + Experience in working cloud services Heroku

, AWS(EC2,S3,Tesseract(OCR), Elastic Beanstalk,[SageMaker](https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor.html)),

GCP(Auto-ML, Google Data Studio, Google Big Query) and Azure(Azure Pinelines, ML Studio)

* + Can handle multiple projects simultaneously

Additionally, I have expertise in developing impactful dashboards using tools like Power BI, Tableau, Google Data Studio, and the ELK Stack

→ Flowise, Voiceflow, Make.com (workflow)

Data Analysis - made easy with LLM

Worked on Large Datasets (10 TB of data)

RDBMS: Google Big Query, DataLake(Snowflake) NoSQL(MongoDB), MySQL, Oracle, MSSQL,

In ML/NLP I have used many libraries

**SpaCy**: SpaCy is an open-source NLP library which is used for Data Extraction, Data Analysis, Sentiment

**Natural language Toolkit (NLTK)**: NLTK is a complete toolkit for all NLP techniques.

**Gensim**: Gensim works with large datasets and processes data streams.

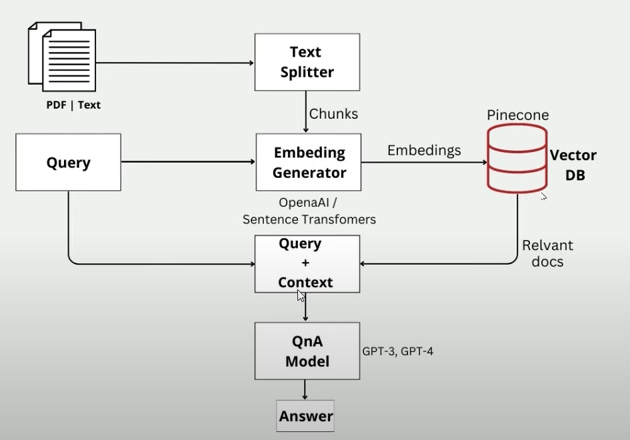
LangChain, Prompt Engineering,

**We will use LLM’s reasoning capability but not knowledge**

Also worked many a times with Fine Tuning(BERT(bert-base-uncased))

Scikit-learn, tensorflow, hugging face, keras, OpenAI,Large Language Model(LLM, BERT, GPT3.5 Turbo, **WizardLM, Huggingface,LLaMA,GooglePalm, T5 flan\_20B,flan-t5-xxl,gpt-j-6B**),Generative AI,

Information Extraction (REG\_EX, Web Scraping, NLP Preprocessing, LLM-Prompt Engineering)



PineCone , StreamleatCHAT(st-chat) Chat Library for UI

**Vector Storage** — CromaDB, Redish, AtlasDB(Indexing storage solutions)

Also used many NLP’s text to number conversion techniques

1. Label encoder
2. One hot encoder
3. Count vectorizer
4. TFIDF vectorizer
5. **N-Gram Vector - fasttext**
6. **Co-Occurrence Matrix - glove**
7. **Word2vec**
8. **Glove**
9. **Fasttext**
10. **BERT**
11. **GPT**

**NLP Pipeline**

**token.lemma\_**

**token.pos\_**

**token.label\_**

**token.is\_stop**

**token.is\_alpha**

**1. Sentence Segmentation :** breaks the paragraph into separate sentences.

**2. Word Tokenization :** break the sentence into separate words or tokens.

**3. Stemming :** Stemming is used to normalize words into its base form or root form. For example, celebrates, celebrated and celebrating, all these words are originated with a single root word "celebrate." The big problem with stemming is that sometimes it produces the root word which may not have any meaning.

**For Example**,

intelligence, intelligent, and intelligently, all these words are originated with a single root word "intelligen." In English, the word "intelligen" do not have any meaning.

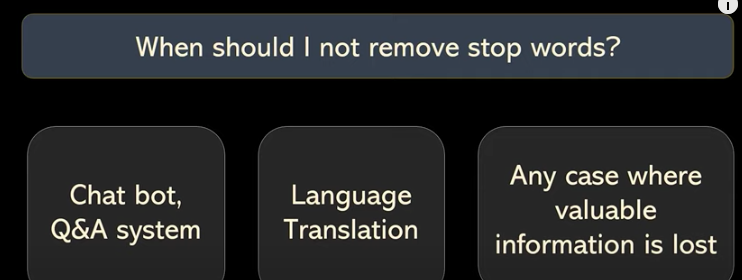
**4. Lemmatization**

Lemmatization is quite similar to the Stamming. It is used to group different inflected forms of the word, called Lemma. The main difference between Stemming and lemmatization is that it produces the root word, which has a meaning.

**For example**: In lemmatization, the words intelligence, intelligent, and intelligently has a root word intelligent, which has a meaning.

**5. Identifying Stop Words**

In English, there are a lot of words that appear very frequently like "is", "and", "the", and "a". NLP pipelines will flag these words as stop words. Stop words might be filtered out before doing any statistical analysis.



**6. POS tags**

POS stands for parts of speech, which includes Noun, verb, adverb, and Adjective. It indicates that how a word functions with its meaning as well as grammatically within the sentences. A word has one or more parts of speech based on the context in which it is used.

**7. Named Entity Recognition (NER)**

Named Entity Recognition (NER) is the process of detecting the named entity such as person name, movie name, organization name, or location.

I have worked in many NLP projects

**1. Auto Pilot (Generative AI Application for DataAnalysis(QnA, Key Insights, Auto Generated Charts(microsoft’s lida library, pandas profiler, ), auto suggest next question based on previous questions))**

* **1. QueryBot: A Natural Language Query System for Customer Support System**QueryBot is an advanced natural language query system where customer can ask any query to the bot and bot will understand the customers query and will answer appropriate answer.  
     
  Client : NOON.COM - E-commerce Platform  
   Technology : NLP, LLM, GPT3.5, LangChain, Prompt Engineering, Python, streamlit  
   **2. QueryBot: A Natural Language Query System for Databases**QueryBot is an advanced natural language query system that enables seamless interaction with databases, allowing users to ask questions in plain language and receive accurate results.  
     
  Client : NOON.COM - E-commerce Platform  
   Technology : NLP, LLM, GPT3.5, LangChain, Prompt Engineering, Python, streamlit **3. ChatBot: A Natural Language Query System for CSV,Excel, PDF Files**ChatBot for Multiple CSV, Excel and PDF file so that you can query on files and get the results Client : NOON.COM - E-commerce Platform  
   Technology : NLP, LLM, GPT3.5, LangChain, Prompt Engineering, Python, streamlit

**1. Semantic Product Search Module**

Improved product search results by using NLP and semantic search techniques to eliminate 100% of faulty products to be displayed

Eg. Vanilla Chocolate → but in a result i can see some vanilla ice cream too.

Italian, arabic and english

Client : NOON.COM - E-commerce Platform

Technology : NLP | BERT | Machine learning | Python

**2. Web Log Analysis for NOON.COM**

Conducted web log analysis on a monthly, daily, and weekly basis.

With extracting information such as visitor counts, peak operating hours, trending links, counts of paid clicks, counts of one-time visitors, counts of returning and non-returning customers, frequency of visits by individual users, and identification of bot visitors.

Log file → powerbi charts

Client : NOON.COM - E-commerce Platform

Technology : JSON | EDA | Python

**3. Implemented NLP-based product recommendation engine**

By analyzing customer purchase history, logs, reviews, feedback, and preferences

Cosine similarity - by calculating euclidean distance of vector

**4. Regression Analysis between content quality score(CQS) and unit conversion rate(UCR)**

A strong correlation was discovered through regression analysis between content quality score (CQS) and unit conversion rate (UCR), indicating the impact of content quality on conversion rates

**5. Customer Segmentation Analysis**

I used the KMeans algorithm to perform Customer Segmentation Analysis and group customers into distinct clusters based on their shared characteristics and behaviors. The analysis included Univariate, Bivariate, and Multivariate approaches.

covering various segmentation methods such as demographic, behavioral, geographic, needs-based, and value-based segmentation etc.

some of the cluster is hidden that too we can identify and visualize

Client : NOON.COM - E-commerce platform

Technology : KMeans Algorithm| Machine learning | Python | EDA

**7. Auto Un-employement Claim Processing** (US Government)

Developed an NLP-based module for automated processing of unemployment claims, capable of extracting information from scanned PDFs and determining claim acceptance or rejection.

Client : US4U

Technology : NLP, Deep Learning, BERT, Python, Amazon Textract OCR

**8. Sonography Image Analysis and Object Detection – Computer Vision**

Developed deep learning-based computer vision convolutional neural network (CNN) model for analyzing and detecting objects in sonography images

Client : Pharmarack

Technology : Computer Vision, CNN, Deep Learning

**9. Sambhavna : Retailer level sell and stock predction Module**

Created a predictive dashboard for retailers to forecast stock availability and sales.

Client : Pharmarack

Technology : Python, Flask, SQL, Machine Learning, AWS

10. **Extrapolation : pincode level aggregation and estimation**

**-> pincode level insights (GSK->All Distributors->medicals)**

**10. Route Optimization Logistic Module and Truck Loading Algorithm for Pharmarack Supply Chain**

Developed a logistics module for route optimization, order distribution, and truck loading algorithm in the pharmaceutical supply chain.

Client : Pharmarack Supply Chain

Technology : Google AI, Google OR-Tool, Traveling Salesperson Problem, Python based Truck Loading Module

**13. NLP Based Auto Retailer Mapping and Auto Product Mapping Recommendation Engine**

Developed an NLP-based auto retailer mapping system that automatically maps distributors and retailers based on shared characteristics. Additionally, created a product mapping feature for distributors. This implementation has replaced a manual 30-person team dedicated to mapping, resulting in cost reduction and automation.

Client : Pharmarack Supply Chain

Technology : NLP, Text Similarity, Python, SQL

**14. Retailers Feedback sentiment Analysis**

This project aims to implement feedback sentiment analysis as a valuable tool for businesses to extract actionable insights from customer feedback. By leveraging sentiment analysis techniques, the project empowers organizations to make informed decisions, enhance customer satisfaction, effectively manage brand reputation, and gain a competitive edge.

Client : Pharmarack Supply Chain

Technology : NLP, Sentiment Analysis

**15. Time Series**

Predict daily sales for medical stores by analyzing various factors like promotions, competition, holidays, seasonality, and locality. Key activities include data analysis, sales visualization, data treatment, time series modeling, interpreting and refining the model, and evaluating predictions for generating insights.

Client : Pharmarack Supply Chain

Technology : Time Series Analysis, Machine learning, Python

**16. Web Scraping using python**

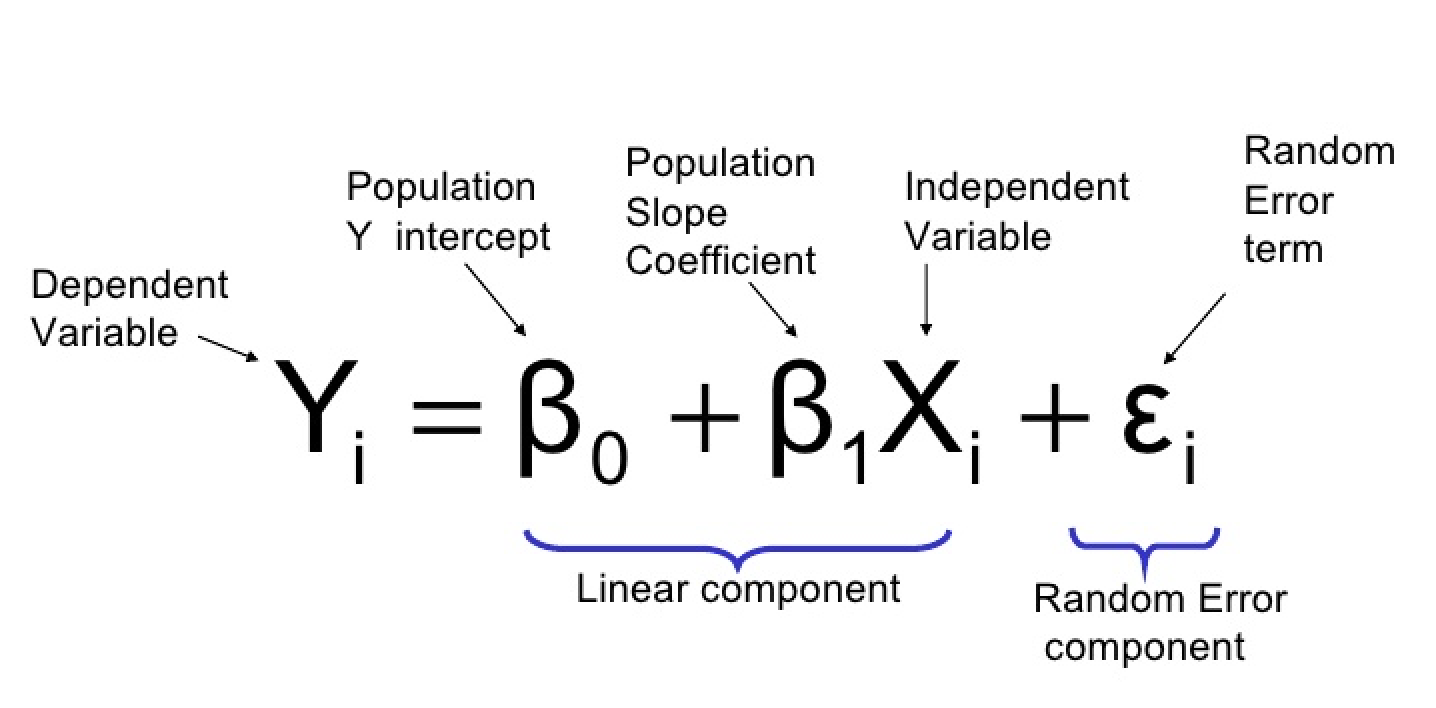
======Statistics

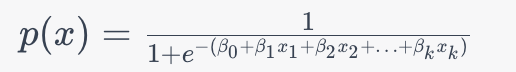
ZScore : A z-score measures exactly how many standard deviations above or below the mean a data point is.

z = data point − mean standard deviation

z = x − μ σ ‍

Regression : to predict continuous numbers



Logistic regression : 

Regression + sigmoid + threshold line(0.5)

Neural N/W -> linear by default so convert it to non-linear using activation function just like logistic regression.

ROC (Receiver Operating Characteristic) Curve: The ROC curve is a graphical plot that shows the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different classification thresholds.

AUC : It ranges from 0 to 1, where a higher AUC-ROC value indicates a better-performing model.

**Difference between R-square and Adjusted R-square**

1. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines. Whereas Adjusted R-squared increases only when independent variable is significant and affects dependent variable.
2. In the table below, adjusted r-squared is maximum when we included two variables. It declines when third variable is added. Whereas r-squared increases when we included third variable. It means third variable is insignificant to the model.

| Macintosh:private:var:folders:h5:q4fq1hm9721clq71yk7_d6zm0000gn:T:TemporaryItems:image_rsq.pnghttps://1.bp.blogspot.com/-ixZb6IYiS6c/U_4DRNokxnI/AAAAAAAADPc/hDelIMR8yEs/s1600/image_rsq.png |
| --- |
| R-Squared vs. Adjusted R-Squared |

Recommendation system

What user want → User Preference → Suggest

4 Type

1. Popularity based recommendation system
   1. Trending Products(sells, rating) formula based
2. Content based recommendation system
   1. Similar product based on same characteristics (attributes)
3. Collaborative filtering based
   1. In user-based collaborative filtering, recommendations are made to a user based on the preferences and behaviors of users who are similar to them.
   2. By finding euclidean distance
4. Hybrid Recommendation system
   1. Combination of all

* Profession/Role: Data Scientist
* Key Responsibilities: Data analysis, predictive modeling, data visualization.
* Knowledge or Expertise: Statistics, machine learning, data wrangling.
* Typical Challenges: Data quality, algorithmic bias, scalability.
* Current Projects: Customer segmentation, predictive maintenance models
* Jargon or Terminology: Regression, clustering, deep learning, ETL.
* Goals and Objectives: Actionable insights, data-driven decision-making.
* Interactions: Business analysts, data engineers, executives.
* Tone and Formality: Analytical, insightful, data-centric.
* Level of Detail: Detailed analysis methods, algorithm explanations.
* Preferred References: Scientific papers, data science frameworks.
* Examples or Analogies: Successful data science projects, industry applications.
* Avoidance of Ambiguity: Clear analytical findings and model explanations.
* Resource Links: Data science libraries, research papers, online courses.
* Follow Up Questions: Query specific data problems, preferred analysis techniques.
* Tables: Data summaries, model performance metrics. Only when necessary.
* Problem Solving Method: Methodical data analysis, model selection, and validation steps.

This data represents time series price data and I want to eventually find trends with this data. first i need you to clean and format this data in order to get it prepare for the analysis