DATA SCIENCE JOB SALARIES

DATA MINING

INTRODUCTION

The majority of information technology students are curious about knowing the salary levels in the field of data science. We will build a model for a data set of employees in the field of data science in different companies and study this data to reach results that help visualize the salaries of these employees and what are the factors affecting them.

PROBLEM

What is the salary range for data science professionals in different companies?

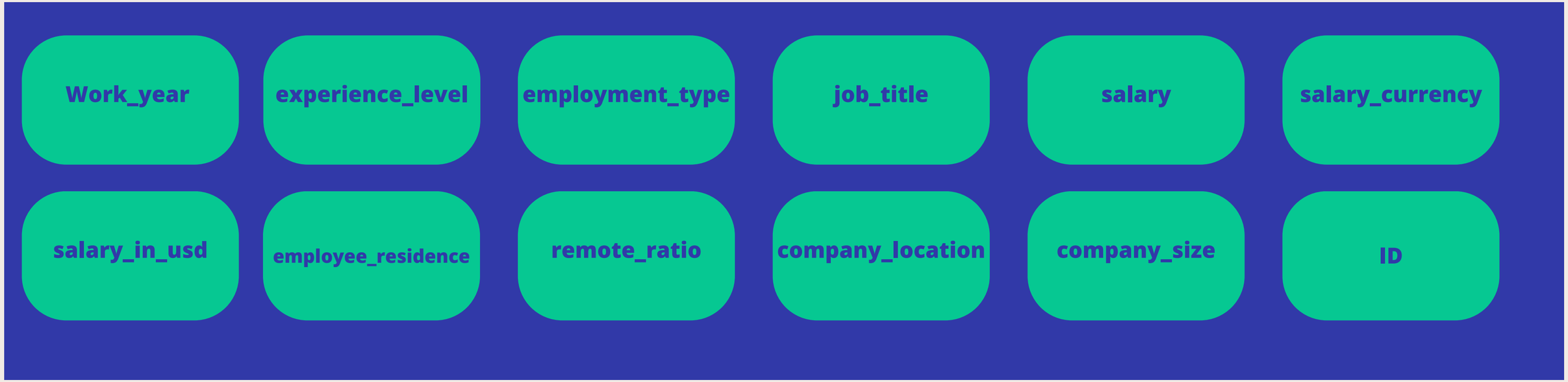
What are the key factors affecting salaries in data science? We will examine various factors such as job title , location, and company size to identify their impact on salaries.

DATA [1]

We applied our data mining tasks on data set consisting of:

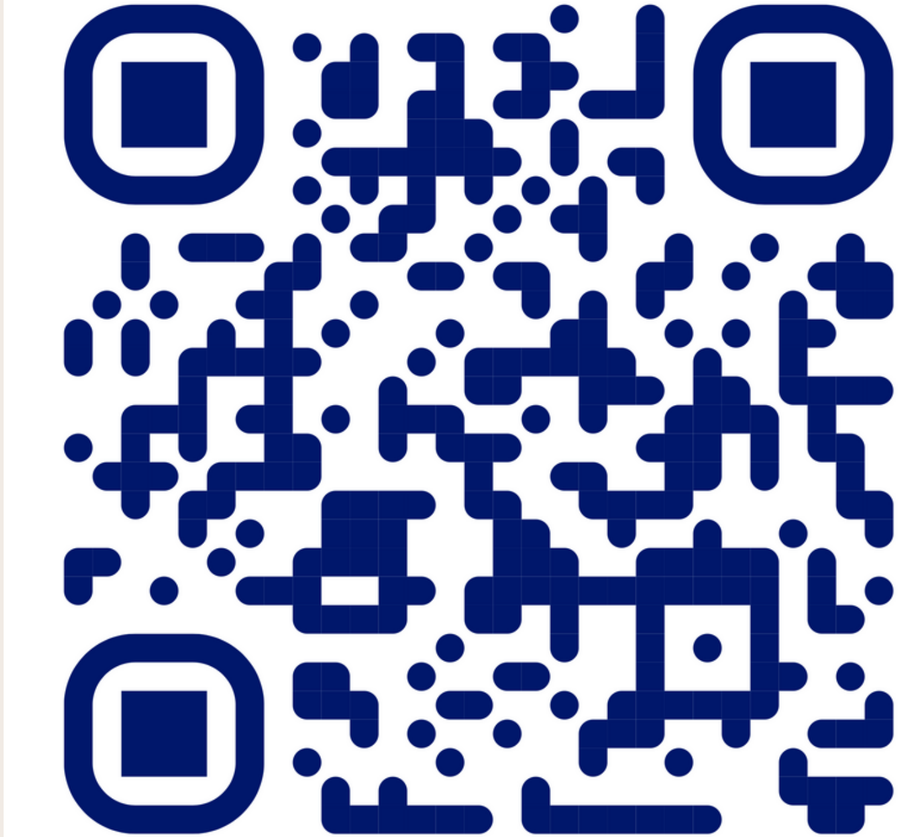
columns: 12

rows: 607



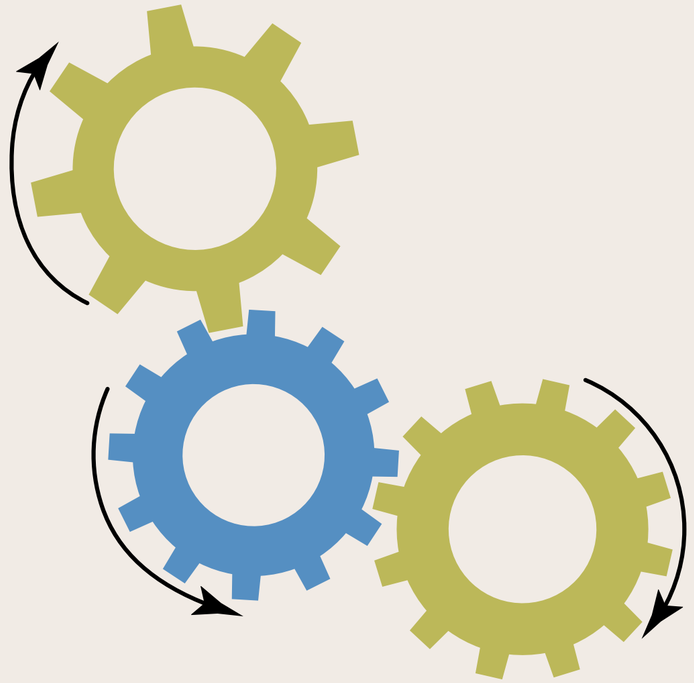
Data source : Data science job salaries

SCAN ME



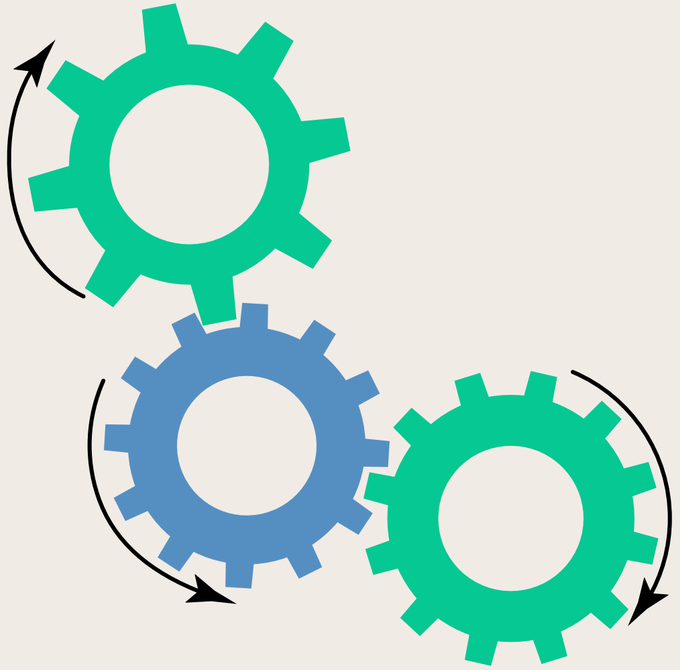
DATA PREPROCESSING

Data preprocessing is a crucial step in the data analysis. It involves transforming raw data into a format that is suitable for analysis and modeling.



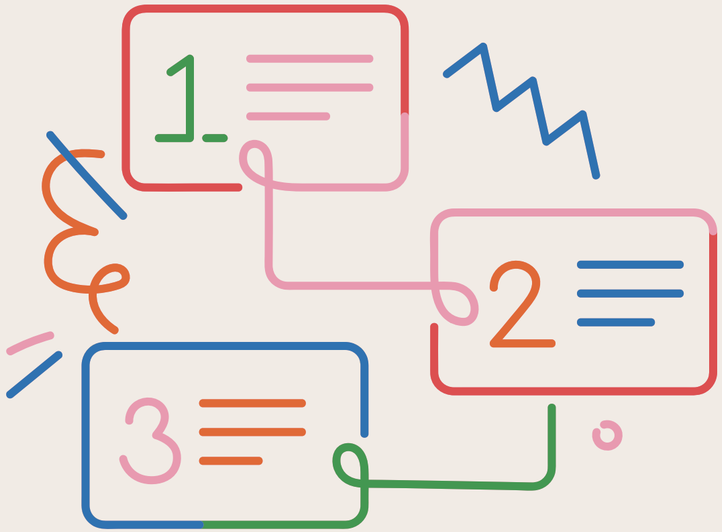
DATA PREPROCESSING

- Encoding
- Dimensionality Reduction
- Outliers Checking
- Normalization
- Missing Values Checking



DATA MINING TASK

In our project we used two data mining tasks to help us predict the data science salary in usd :



1. Regression

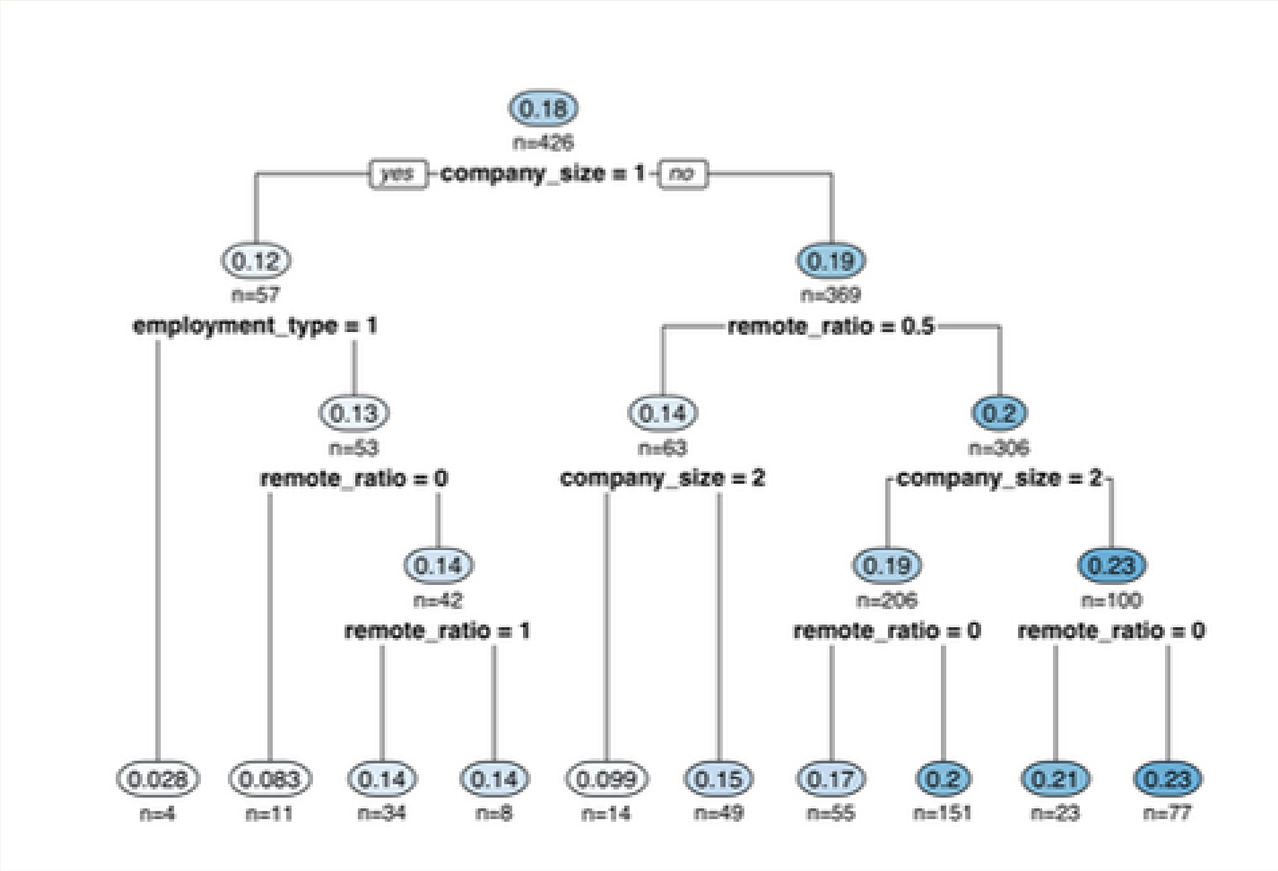
2. Cluestring

REGRESSION [2]

We built tow decision tree models in regression step :

- the first tree model uses three splitting criterion used in the decision tree algorithm "information ", "gain", and "gini" and then we measure the RMSE of each criterion, which is a measure of the average deviation between the predicted values of the model and the actual values in the dataset.

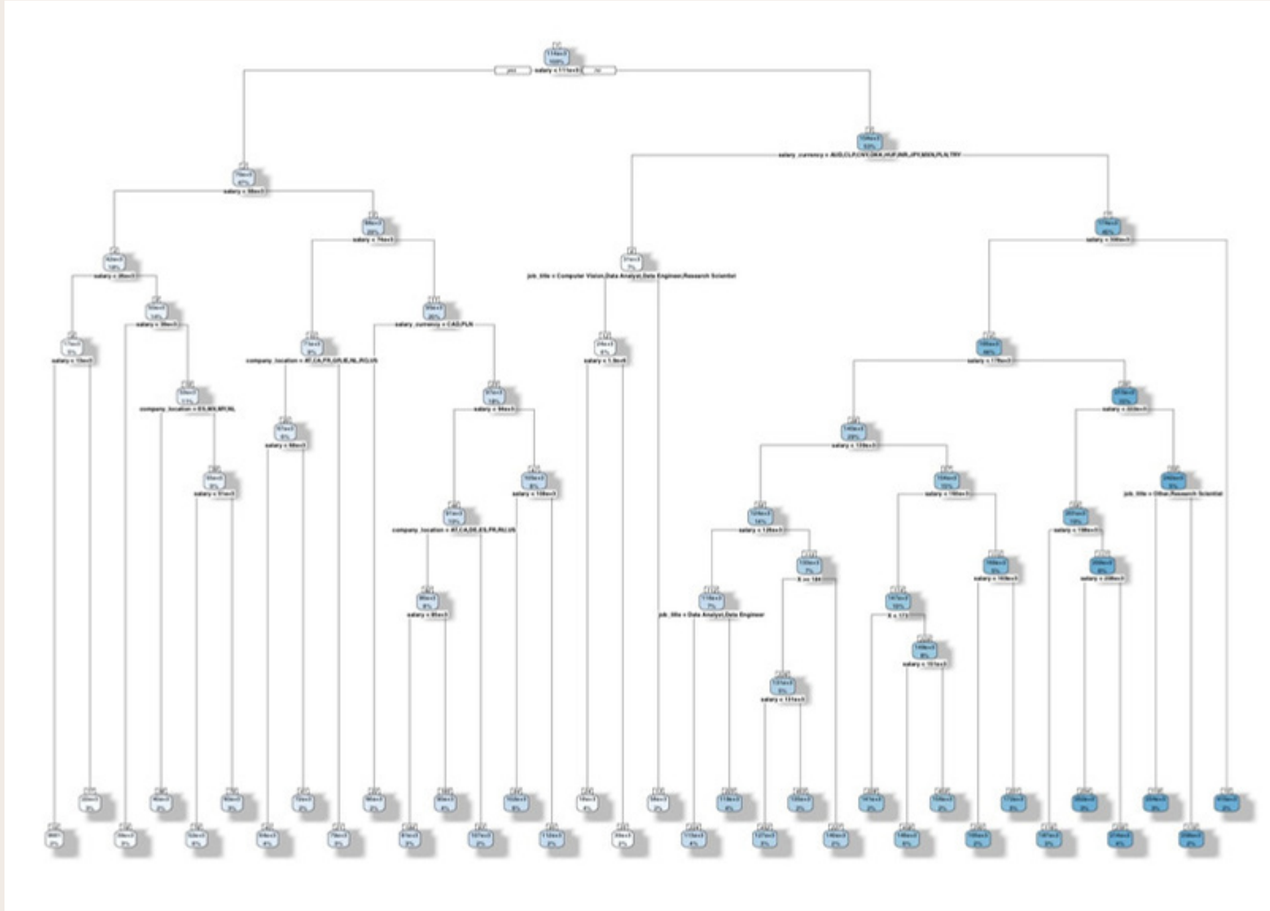
- the optimal choice of tree model to select Model with 70% of training set and 30% of test set , This decision is guided by the fact that Model exhibits the most favorable RMSE value .



REGRESSION

- the second tree model uses three splitting criterion used in the decision tree algorithm "anova", "gini", and "dev" and then we measure the accuracy of each criterion, and We set the tolerance value to be 1000 tolerance value implies that the model's predictions will be considered accurate if the absolute difference between the predicted salary and the actual salary is within 10000 unitsif the absolute difference exceeds 10000 units, the prediction will beconsidered inaccurate.

- the optimal choice of tree model to select Model with 70% of training set and 30% of test set too , This decision is guided by the fact that Model exhibits the most favorable accuracy value .



CLUESTRING

Choosing Best K Value for K-means Clustering

the first clustering algorithm we've selected the attributes that we find more interesting in our study to see what the best K mean on these attributes which is ("work_year", "salary_in_usd", "remote_ratio", "experience_level", "employment_type", "company_size").

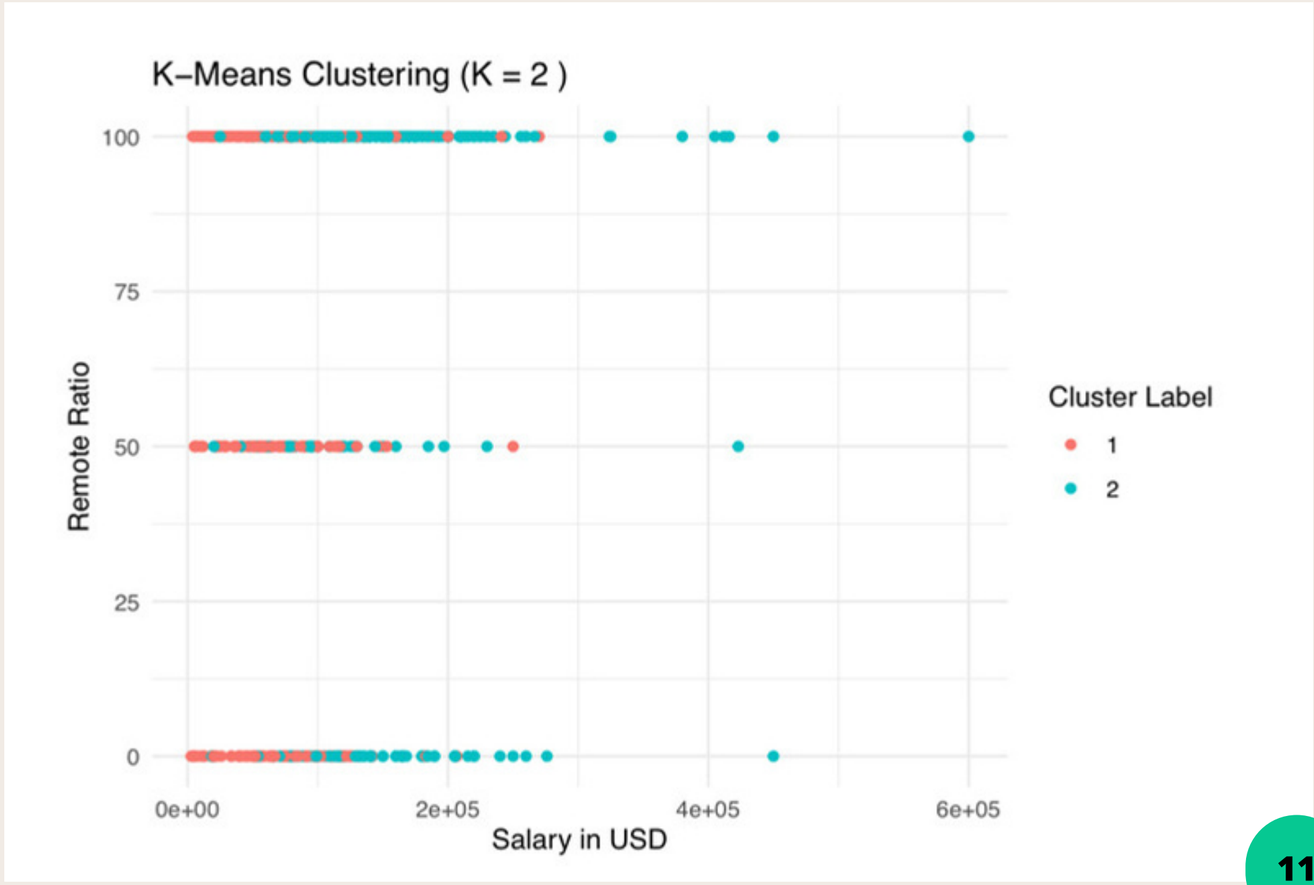
the result for the selected attributes:

Silhouette Scores:

- K = 2: Silhouette score = 1
- K = 3: Silhouette score = 3
- K = 4: Silhouette score = 3

Total WSS (Within-Cluster Sum of Squares):

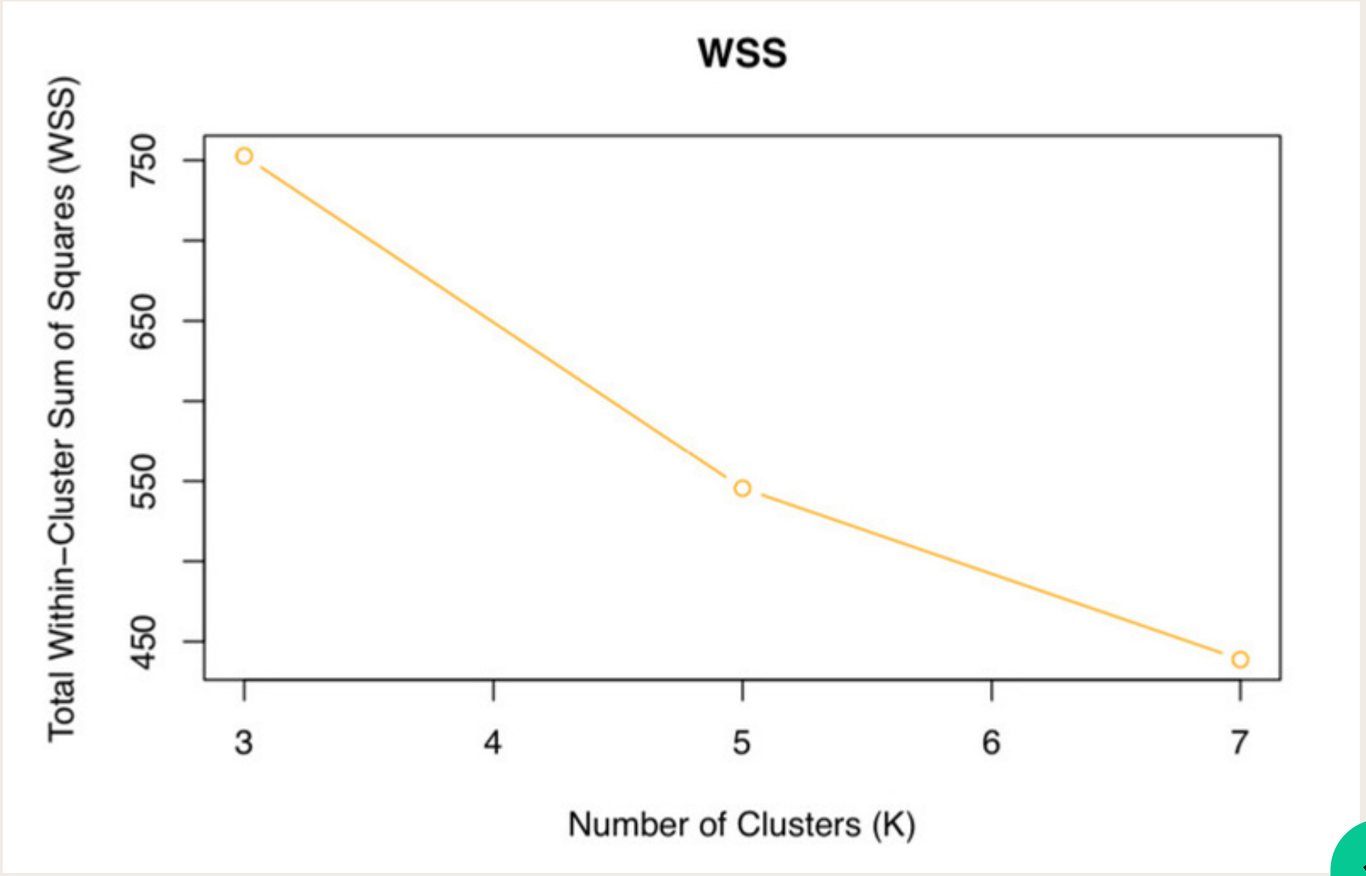
- K = 2: WSS = 6035.376
- K = 3: WSS = 5106.642
- K = 4: WSS = 4246.153



CLUESTRING

Choosing Best K Value for K-means Clustering

A good K value typically strikes a balance between a high Silhouette score and a low WSS, in the second clustering algorithm we did in our dataset indicating that K = 7 has the highest Silhouette score which indicates good cluster separation. Additionally, K = 7 has the lowest WSS indicating tight and compact clusters, K = 7 appears to be a good choice.



Silhouette Scores:

- K = 3: Silhouette score = 0.3519878
- K = 5: Silhouette score = 0.3217594
- K = 7: Silhouette score = 0.4063718

Total WSS (Within-Cluster Sum of Squares):

- K = 3: WSS = 758.8789
- K = 5: WSS = 652.4071
- K = 7: WSS = 444.8279

FINDING

Our model uses a regression approach with a 70-30 data split and employs the RMSE method for accurate predictions. The use of regression instead of classification ensures better handling of continuous salary values. The model aims to provide valuable insights for individuals making career decisions and assists companies in establishing fair pay structures. The focus on simplicity and accuracy makes understanding data science salaries accessible to all stakeholders.

REFERENCES:

[1] HTTPS://WWW.KAGGLE.COM/DATASETS/RUCHI798/DATA-SCIENCE-JOB-SALARIES

[2] HTTPS://MEDIUM.COM/NERD-FOR- TECH/IMPLEMENTING-DECISION-TREES-IN-R-REGRESSION-PROBLEM-USING-RPART-C74CBD9E0B7B

THANK YOU FOR LISTENING

RUBA ALBNHAR – REMA ALSHAREF – MARYAM ALRAWAYAH