




# REPORT 2

# PREDICTIVE ANALYTICS

Prateek Majumder  
prateekmaj21@gmail.com



## The Data

The dataset contains 12043 transactions for 100 customers who have one bank account each. Transactional period is from 01/08/2018 - 31/10/2018 (92 days duration). The data entries are unique and have consistent formats for analysis. For each record/row, information is complete for majority of columns. Some columns contain missing data (blank or NA cells), which is likely due to the nature of transaction. (i.e. merchants are not involved for Inter Bank transfers or Salary payments) It is also noticed that there is only 91 unique dates in the dataset, suggesting the transaction records for one day are missing (turned out to be 2018-08-16).

The columns are-

```
['status', 'card_present_flag', 'bpay_biller_code', 'account',  
  'currency', 'long_lat', 'txn_description', 'merchant_id',  
  'merchant_code', 'first_name', 'balance', 'date', 'gender', 'age',  
  'merchant_suburb', 'merchant_state', 'extraction', 'amount',  
  'transaction_id', 'country', 'customer_id', 'merchant_long_lat',  
  'movement']
```

- The data has numeric features, but they are few. Useful numeric features include the customers' age, transaction amount and bank balance.
- We extract out the customers' salary from the transaction description column and get the amount for that row, giving us the salary.
- We take average (mean) values for salary, amount, balance and age[which gets no mean].
- We make scatterplots for- Salary vs Age, Salary vs Account Balance and Salary vs Transaction Amount.
- There is indeed a relation between these features, but the relation is not clear.
- Filtering the type of transaction in the transaction description column, we try to get the amount the person has spent on an average in POS(Point of Sale- Retail, Business, Purchases etc), Payment(Bills etc), Interbank(Bank to Bank transfer) and Phone Bank(probably Internet Banking).
- I see phone bank and bank transfer donot have adequate data for all customers. Hence avoid that data.
- I thus, created 2 new features for the data. The average payments made by customers and the average money spent by customers at point of sale.

- Creating scatterplots for the new features with respect to salary, There seem to be relations, but the relation is not quite clear.
- The main shortcoming is the lack of data. Perhaps with large amounts of data, more relations can be found.

### Linear Regression Model-

Took these columns as X and y. The r2 score was low, main reason being lack of data, hence less training examples.

```
#Defining the X and y of machine learning  
X=df[["age","balance","payment","pos"]].values  
y=df["salary"].values
```

```
#r2 score  
print('Coefficient of determination: ', r2_score(y_test, y_pred))
```

```
Coefficient of determination: 0.07666219584715894
```

But, making a sample prediction (refer python notebook for full details), the model did make predictions in realistic values, of what a person's salary might be. So, with more data, the model could be made better.

## Decision Tree Regressor-

We took the same columns for X and y. This model also gave a very low score. But the sample prediction it made was also in realistic values, of what a person's salary might be. With more training data, the model could be made better.

```
#score using test values  
dt.score(X_test, y_test)
```

```
0.2041445918495219
```

## Customer Segmentation using KMeans-

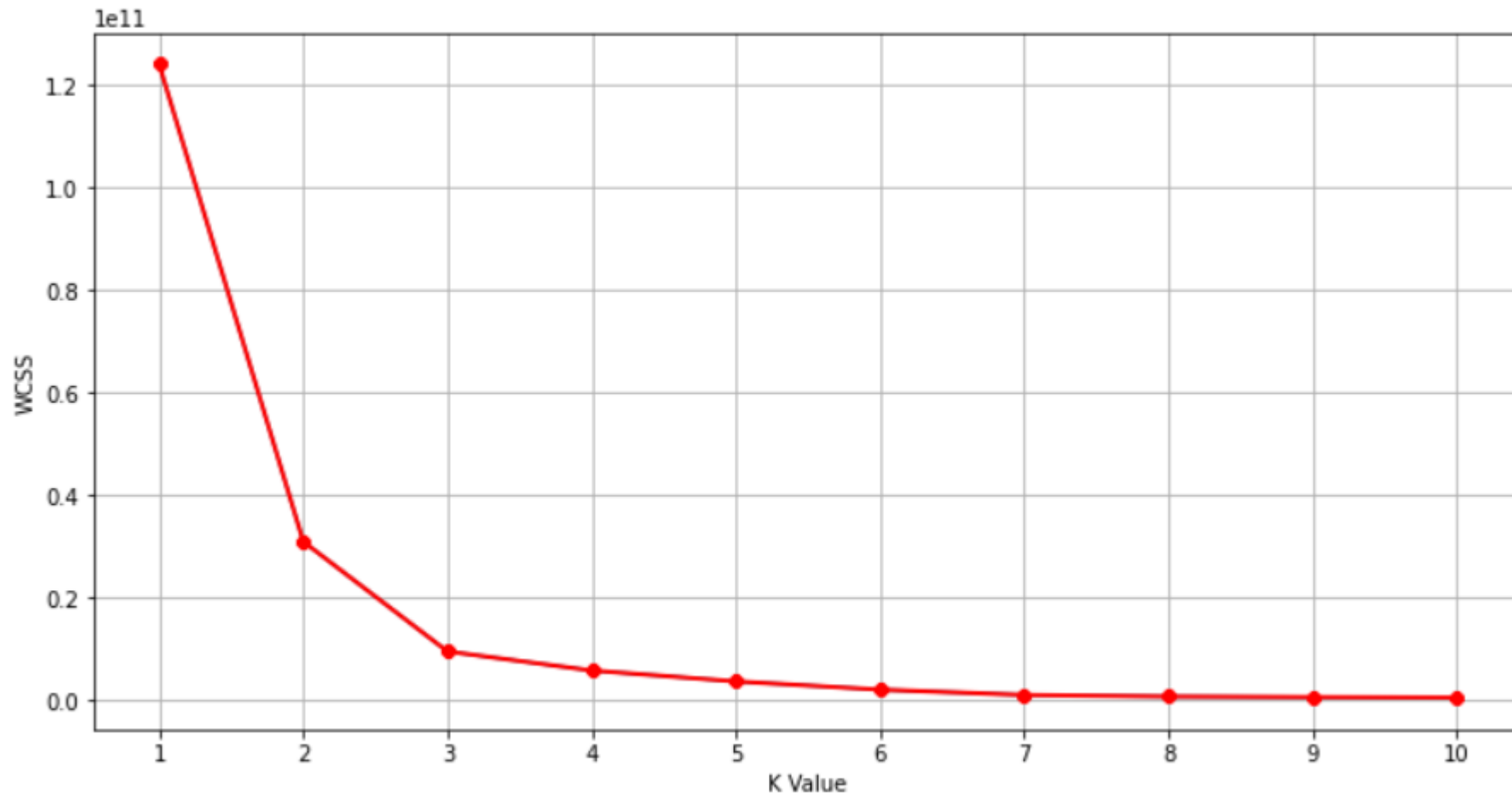
We shall work with an unsupervised learning algorithm to

*#The data we shall be working with*

df

	customer_id	age	balance	payment	pos	salary
0	CUS-1005756958	53	2275.852055	144.000000	28.769615	970.47
1	CUS-1117979751	21	9829.929000	98.925000	23.321923	3578.65
2	CUS-1140341822	28	5699.212250	142.000000	34.781282	1916.51
3	CUS-1147642491	34	9032.841186	96.185185	54.271316	1711.39
4	CUS-1196156254	34	22272.433755	43.100000	31.157432	3903.73
...	...	...	...	...	...	...
95	CUS-72755508	35	4497.557069	1180.000000	25.022143	725.32
96	CUS-809013380	21	3756.902903	70.136364	22.255098	1037.07
97	CUS-860700529	30	3462.276009	41.933333	28.164845	1808.62
98	CUS-880898248	26	8528.830385	77.500000	20.101429	1433.98
99	CUS-883482547	19	9877.452697	91.446809	30.635098	3977.46

KMeans- The elbow curve of Kmeans shows that the elbow is formed at  $K=3$ . So we can make 3 clusters.



KMeans- The customer segments have been made. The segments are based on customer age, account balance, average payment and POS transactions and salary. Please refer to the python notebook for more details.

Final Words- There are correlations between the features. But they are not exactly clear. With more data, we can make more clear assumptions. The regression models are not very accurate. Customer segmentation has been done. The project was a good experience for me to understand analytics in the banking sector.