Exploratory
Data Analysis
- ANZ Transaction Dataset-

ABOUT THE DATASET

Insight into the data

The dataset contains the details of 100 customer's transaction history in Australia from 2018.
The transaction are mostly between the month August to September.
There are total 12043 records & 23 columns.

Null Values

There were quite a lot of null values in dataset. Two columns ['bpay_biller_code', 'merchant_code'] were dropped. The remaining null values were mostly because of 'posted' value in status column. As the transaction details were pending.

Average Transaction amount=187.93358797641767

Maximum Balance amount=267128.52

8 Merchant state =NSW,QLD,WA,SA,NT,ACT,TAS,VIC

Age= range(18 to 78)
Highest count of customer of age=26
Lowest count of customer of age=78

|	status	card_present_flag	bpay_biller	account	currency	long_lat	txn_desc	merchant	merchant	first_nam	balance	date	gender	age	merchant_suburb	merchant_state	extractio	amount	transacti	country	custome	merchant_long_lat	movement
2	authorized	1		ACC-159845107	AUD	153.41 -27.95	POS	81c48296-73be-4		Diana	35.39	8/1/2018	F	26	Ashmore	QLD	2018-08	16.25	a623070	Australia	CUS-248	153.38 -27.99	debit
3	authorized	0		ACC-159845107	AUD	153.41 -27.95	SALES-F	830a451c-316e-4		Diana	21.2	8/1/2018	F	26	Sydney	NSW	2018-08	14.19	13270a2	Australia	CUS-248	151.21 -33.87	debit
4	authorized	1		ACC-12223005	AUD	151.23 -33.94	POS	835c231d-8cdf-4		Michael	5.71	8/1/2018	M	38	Sydney	NSW	2018-08	6.42	feb79e7	Australia	CUS-214	151.21 -33.87	debit
5	authorized	1		ACC-103705051	AUD	153.10 -27.66	SALES-F	48514682-c78a-4		Rhonda	2117.2	8/1/2018	F	40	Buderim	QLD	2018-08	40.9	2698170	Australia	CUS-161	153.05 -26.68	debit
6	authorized	1		ACC-159845107	AUD	153.41 -27.95	SALES-F	b4e02c10-0852-4		Diana	17.95	8/1/2018	F	26	Mermaid Beach	QLD	2018-08	3.25	329adf7	Australia	CUS-248	153.44 -28.06	debit
7	posted			ACC-16083633	AUD	151.22 -33.87	PAYMENT			Robert	1705.4	8/1/2018	M	20			2018-08	163	1005b48	Australia	CUS-2688605418		debit
8	authorized	1		ACC-27762528	AUD	144.95 -37.76	SALES-F	3aa18033-a0a9-4		Kristin	1248.4	8/1/2018	F	43	Kalkallo	VIC	2018-08	61.06	b79ca2C	Australia	CUS-412	144.95 -37.53	debit
9	authorized	1		ACC-27762528	AUD	144.95 -37.76	POS	ee58145d-26e8-4		Kristin	1232.8	8/1/2018	F	43	Melbourne	VIC	2018-08	15.61	e1c4a50	Australia	CUS-412	144.96 -37.81	debit
10	authorized	1		ACC-18244657	AUD	116.06 -32.00	POS	cfbf535e-caa8-4		Tonya	213.16	8/1/2018	F	27	Yokine	WA	2018-08	19.25	799e39e	Australia	CUS-302	115.85 -31.9	debit
11	posted			ACC-60266757	AUD	151.23 -33.96	INTER BANK			Michael	466.58	8/1/2018	M	40			2018-08	21	798a778	Australia	CUS-2031327464		debit
12	posted			ACC-217159328	AUD	146.94 -36.04	PAYMENT			Fernand	4348.5	8/1/2018	M	19			2018-08	27	baff17b2	Australia	CUS-2317998716		debit
13	posted			ACC-27762528	AUD	144.95 -37.76	PAYMENT			Kristin	1203.8	8/1/2018	F	43			2018-08	29	78a1b6c	Australia	CUS-4123612273		debit
14	authorized	1		ACC-18244657	AUD	116.06 -32.00	SALES-F	33952b07-859c-4		Tonya	207.08	8/1/2018	F	27	Cockburn Central	WA	2018-08	6.08	9ba4928	Australia	CUS-302	115.86 -32.13	debit
15	posted			ACC-58856484	AUD	151.27 -33.76	INTER BANK			Isaiah	4438.2	8/1/2018	M	23			2018-08	25	eaafa60	Australia	CUS-1462656821		debit
16	posted			ACC-149645199	AUD	145.16 -37.84	INTER BANK			Ricky	173.66	8/1/2018	M	43			2018-08	39	243dcea	Australia	CUS-3142625864		debit
17	authorized	1		ACC-19030375	AUD	153.05 -27.61	POS	d920de7f-959c-4		Jeffrey	2.85	8/1/2018	M	30	Mount Ommaney	QLD	2018-08	10.79	28347ba	Australia	CUS-860	152.94 -27.55	debit
18	posted			ACC-20148568	AUD	144.99 -37.90	INTER BANK			Patrick	260515	8/1/2018	M	46			2018-08	22	ae8124d	Australia	CUS-2370108457		debit
19	posted			ACC-416382218	AUD	149.03 -34.97	PAYMENT			Karen	3117.9	8/1/2018	F	26			2018-08	55	0b0bc16	Australia	CUS-2630892467		debit
20	posted			ACC-39546778	AUD	115.72 -32.28	PAYMENT			Ruth	38.31	8/1/2018	F	47			2018-08	58	c24ca85	Australia	CUS-3716701010		debit

DASHBOARD

We have a high volume of customer of age between 29 to 40

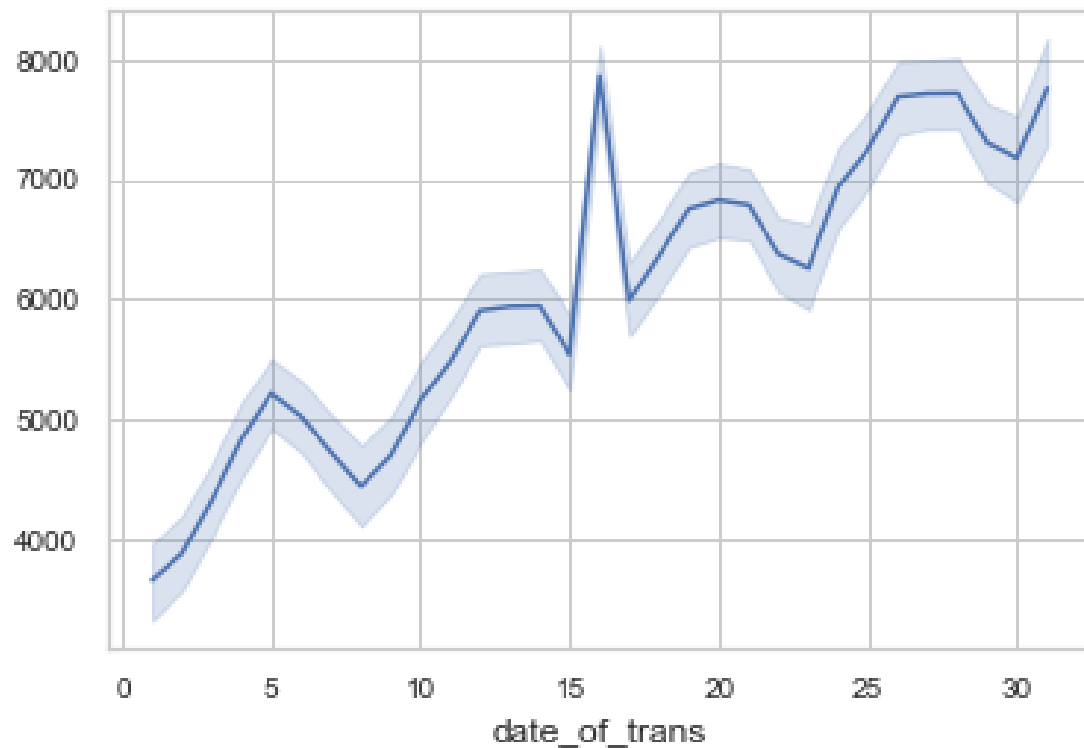Mostly the amount stored is between 0 to 1000 on average on each account.

More customer prefer buying using the card.
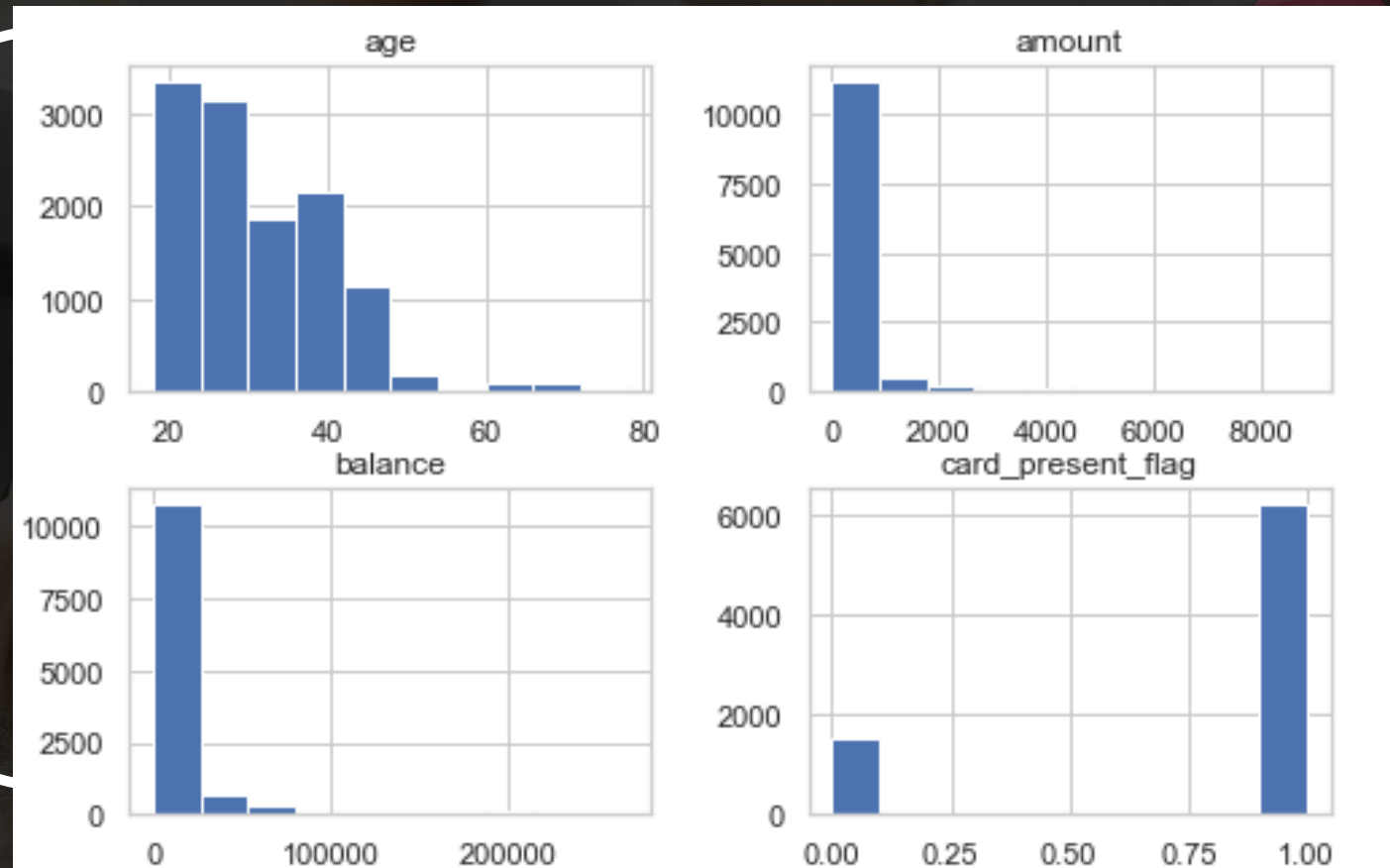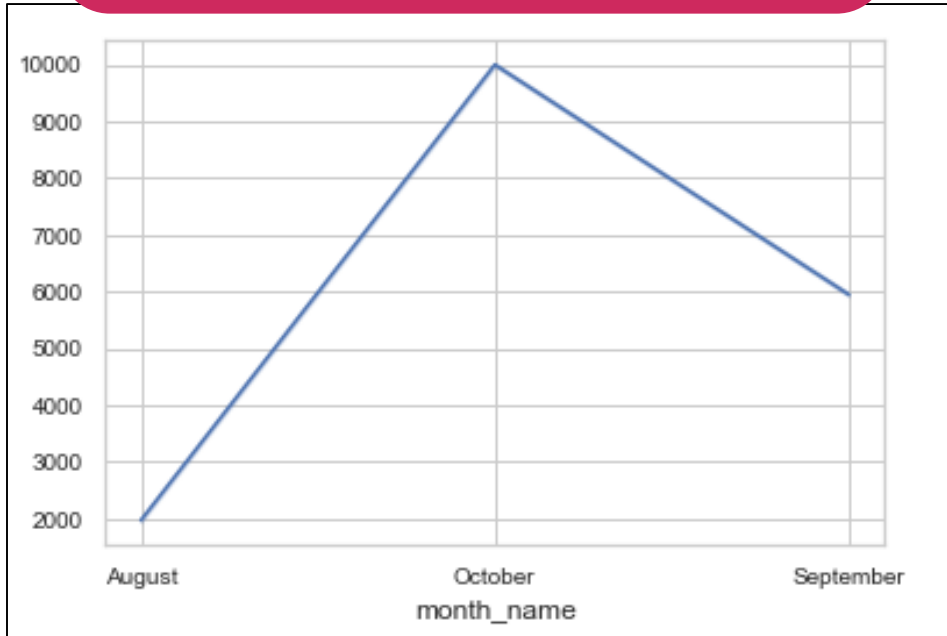
DASHBOARD

Daily Transaction

- ❖ More number of transaction has happened on 17th of the month.
- ❖ The customer with highest transaction is :
  ('CUS-2487424745', 'Diana')
- ❖ The customer with least transaction is:
  ('CUS-1739931018', 'Kimberly')
- ❖ The customer who had the most number of transaction in a day:

(Timestamp('2018-09-14 00:00:00'), 'CUS-860700529', 'Jeffrey')

DASHBOARD

Monthly transaction

Weekly transaction
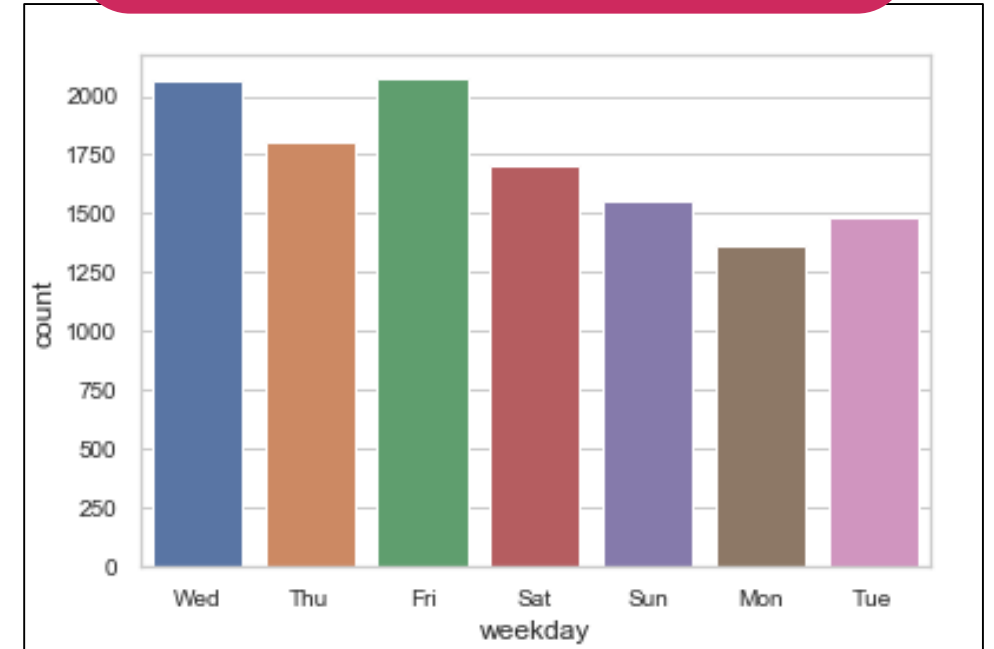
Month	Number of Transactions
August	3943
September	4013
October	4087

Transaction amount increases every day peaking at Friday after which it falls over the weekend

There is a high correlation between age & balance.

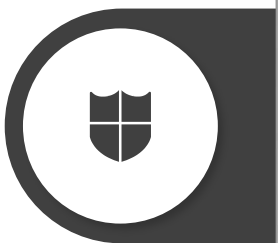The average transaction amount is spend by 52 years of age.

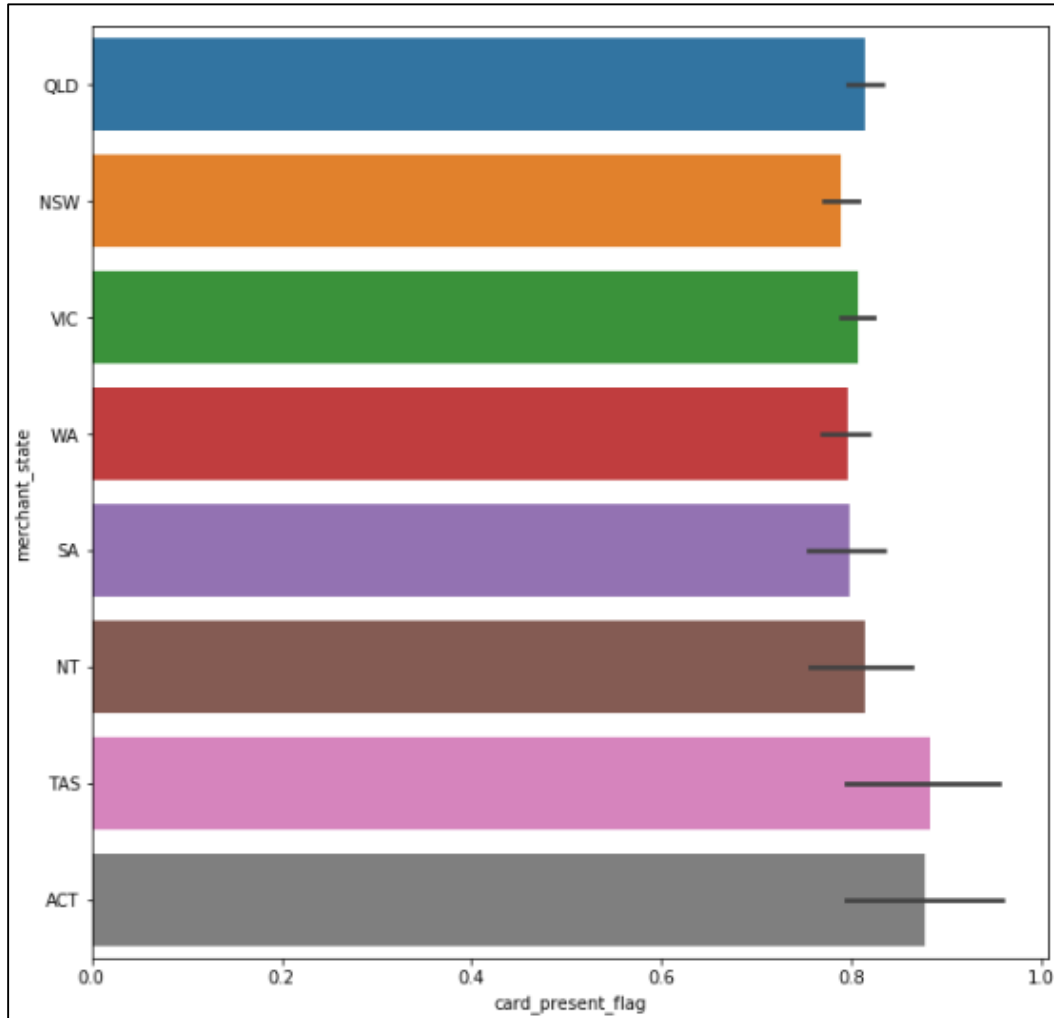There are few outliers who spend a lot in age group 2.
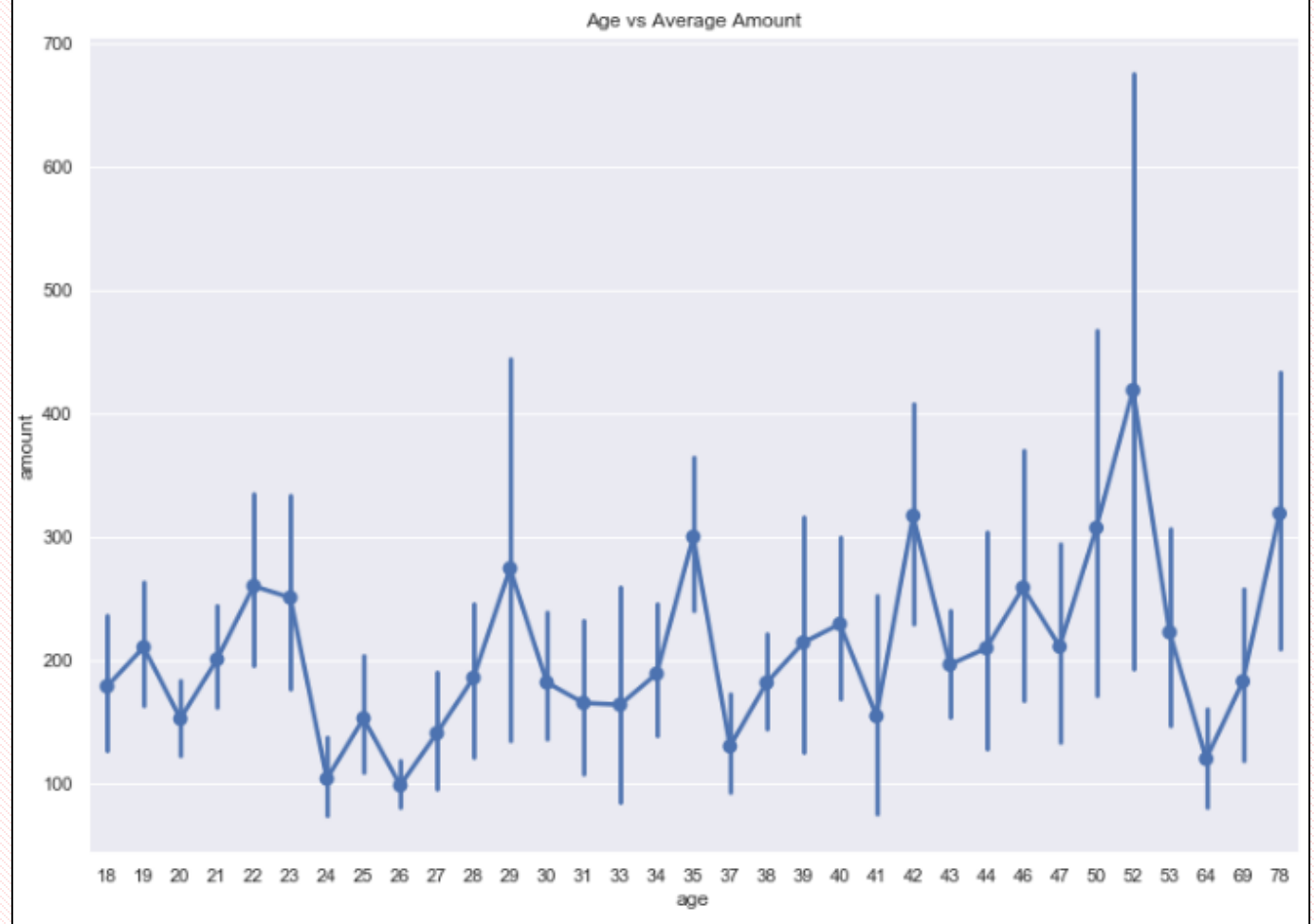
Age vs Average Balance
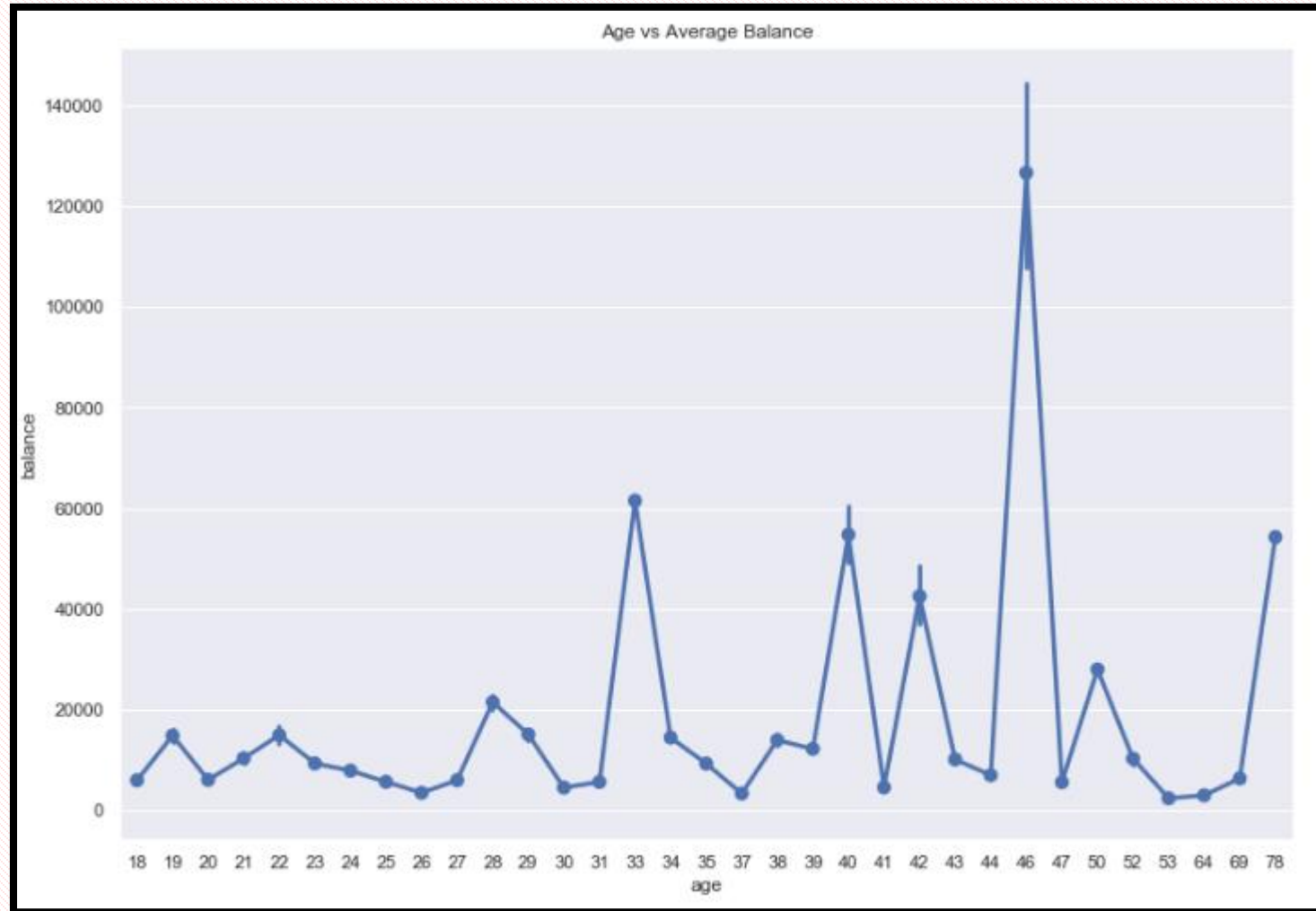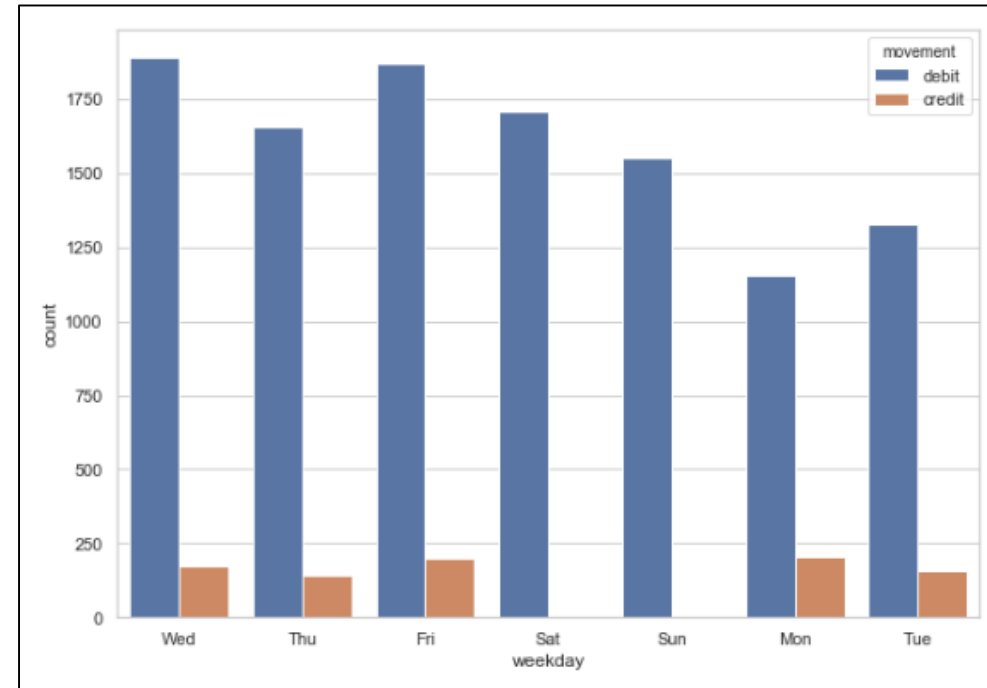
Customers with age 46 years store the highest average balance.
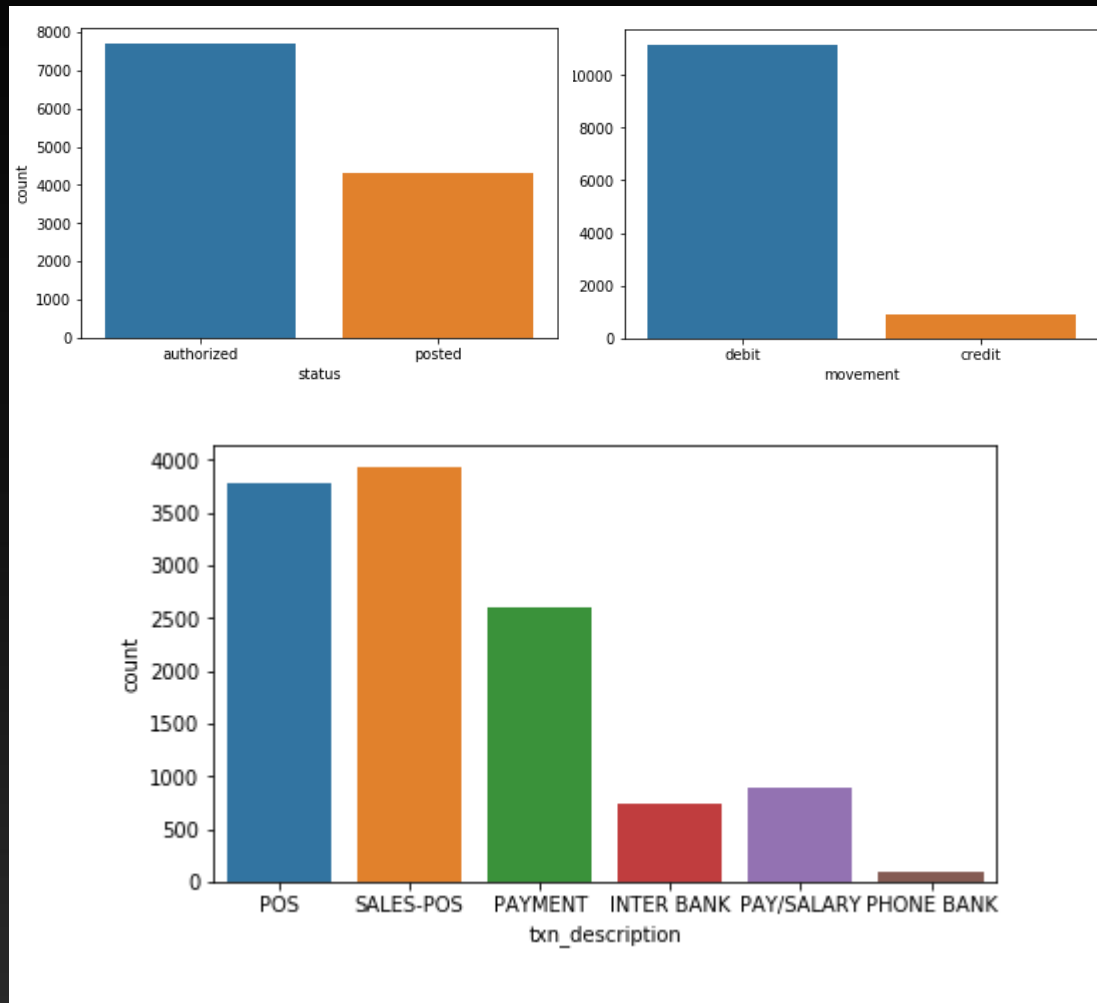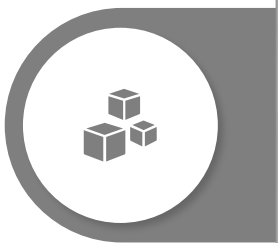
❑ Most transaction has been authorized.
❑ Most customers use debit card.
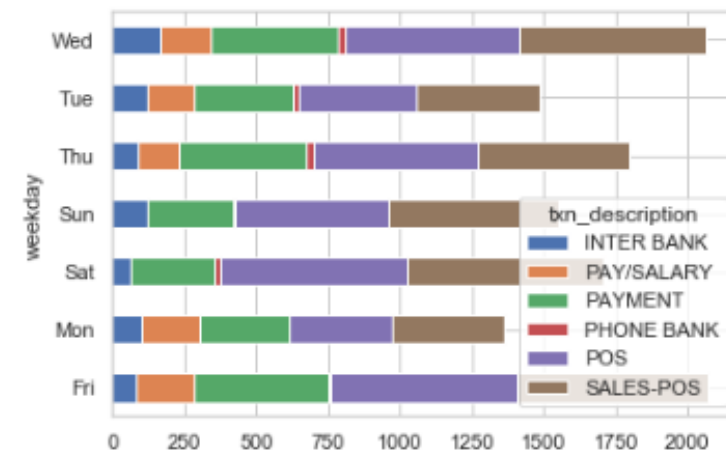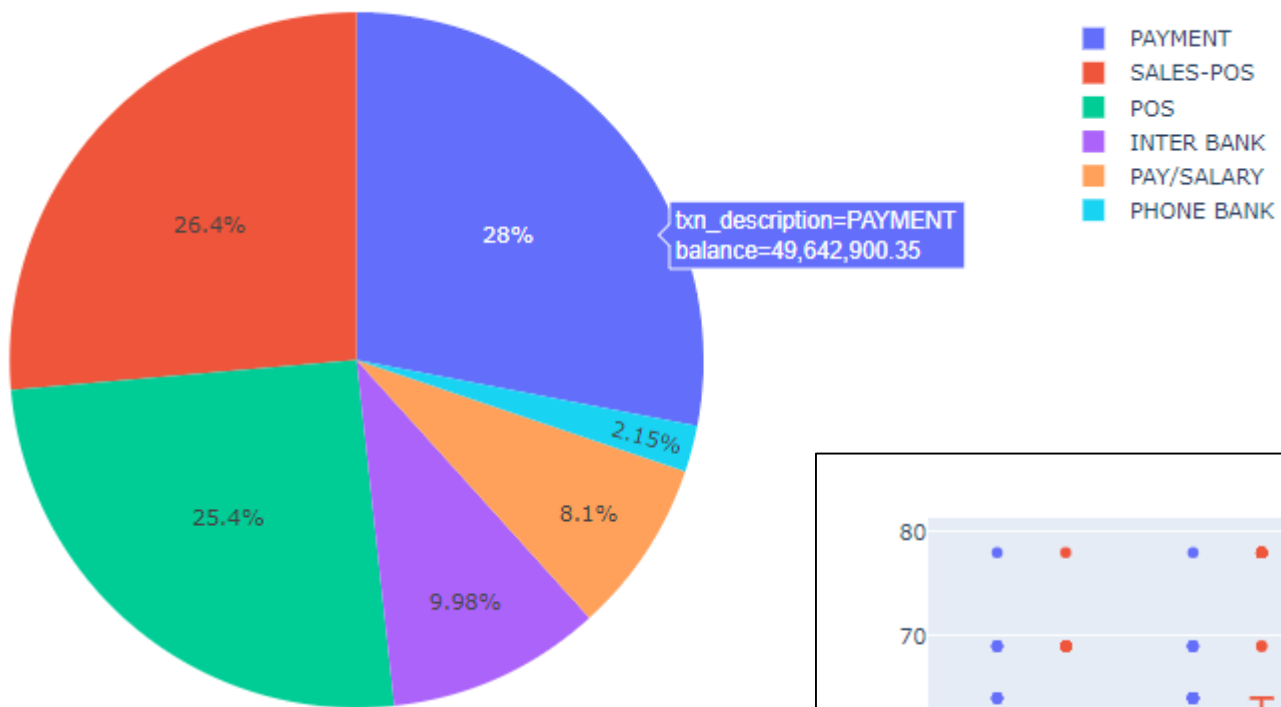❑ Transaction occur increasing from Phone-Bank to Sales-POS

CARD PRESENT % STATE WISE

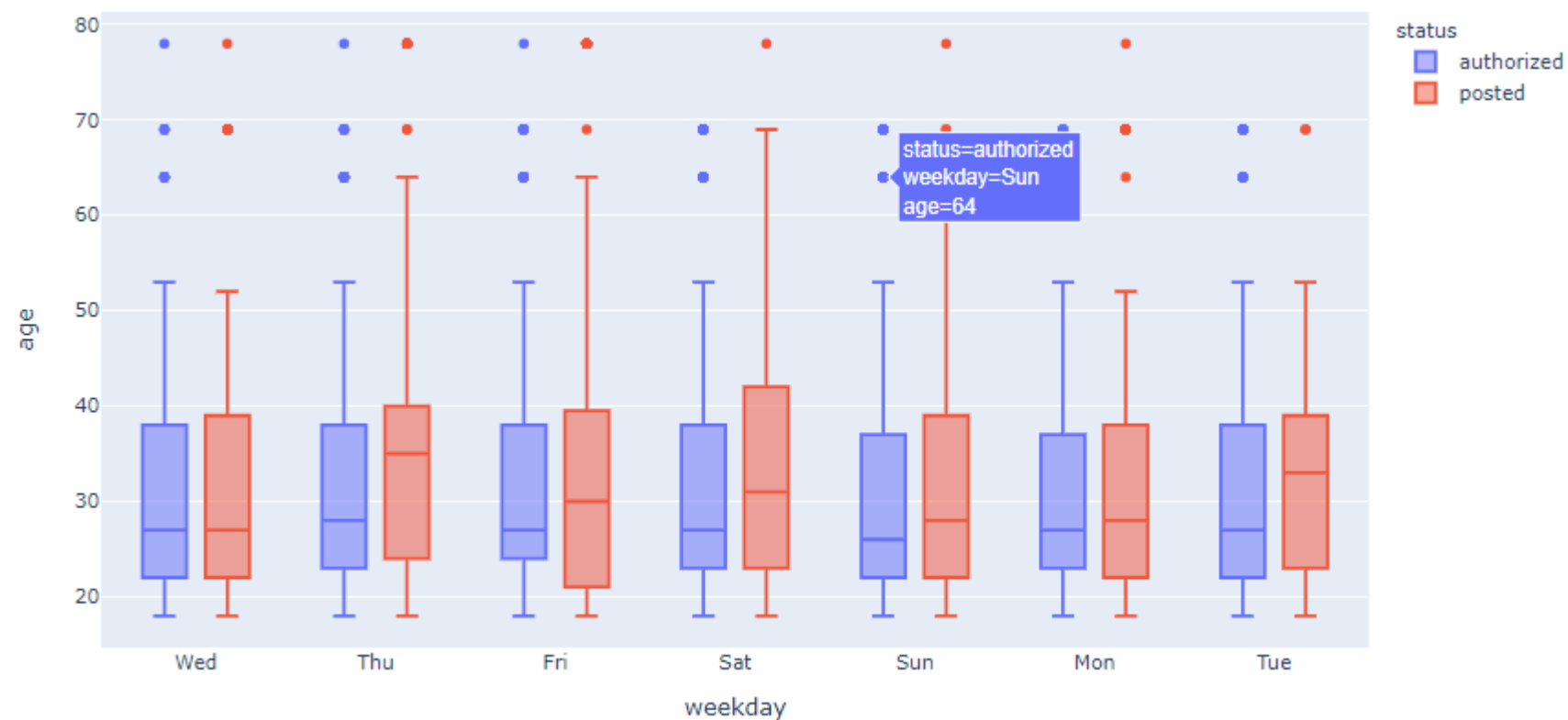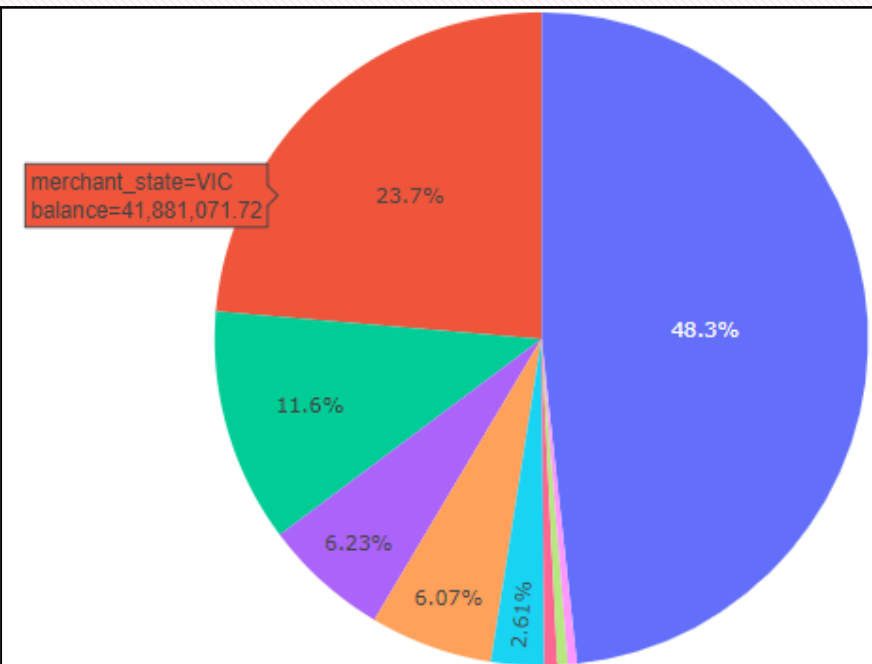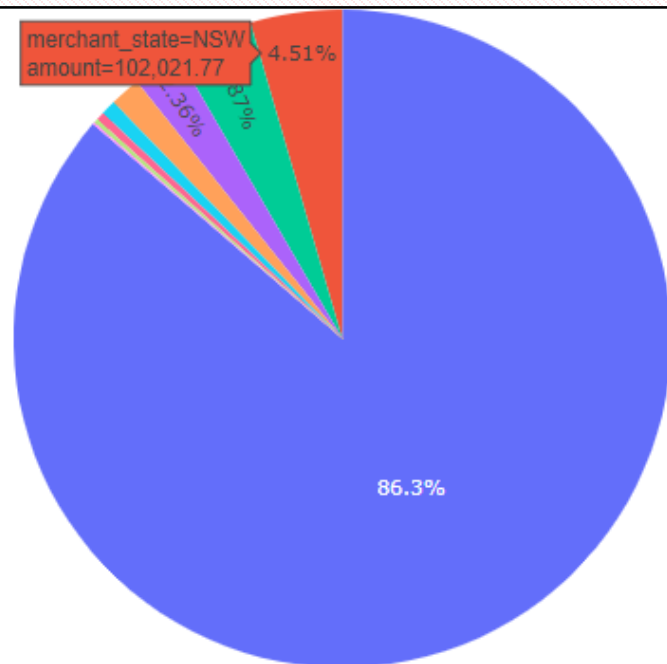*Customers don't buy over the weekend using credit card at all.*

WEEKLY TRANSACTION

Most Transaction between payment, Sales-POS & POS.

An interactive graph that represents transaction weekly done by different age.

STATE WISE HIGHEST TRANSACTION

STATE WISE HIGHEST BALANCE

merchant_state=NSW
amount=102,021.77

4.51%

2.36%

3.7%

86.3%

Legend:
- null
- VIC
- NSW
- QLD
- WA
- SA
- NT
- ACT
- TAS

merchant_state=VIC
balance=41,881,071.72

23.7%

48.3%

11.6%

6.23%

6.07%

2.61%

merchant_state	
NSW	102021.77
VIC	87584.00
QLD	53483.45
WA	33992.06
SA	16776.57
NT	9168.89
ACT	4876.68
TAS	1962.93

merchant_state	
VIC	41881071.72
NSW	20493615.96
QLD	11035157.51
WA	10746485.22
SA	4627386.99
NT	1136194.09
ACT	911719.22
TAS	777312.96

Average amount spent by customer
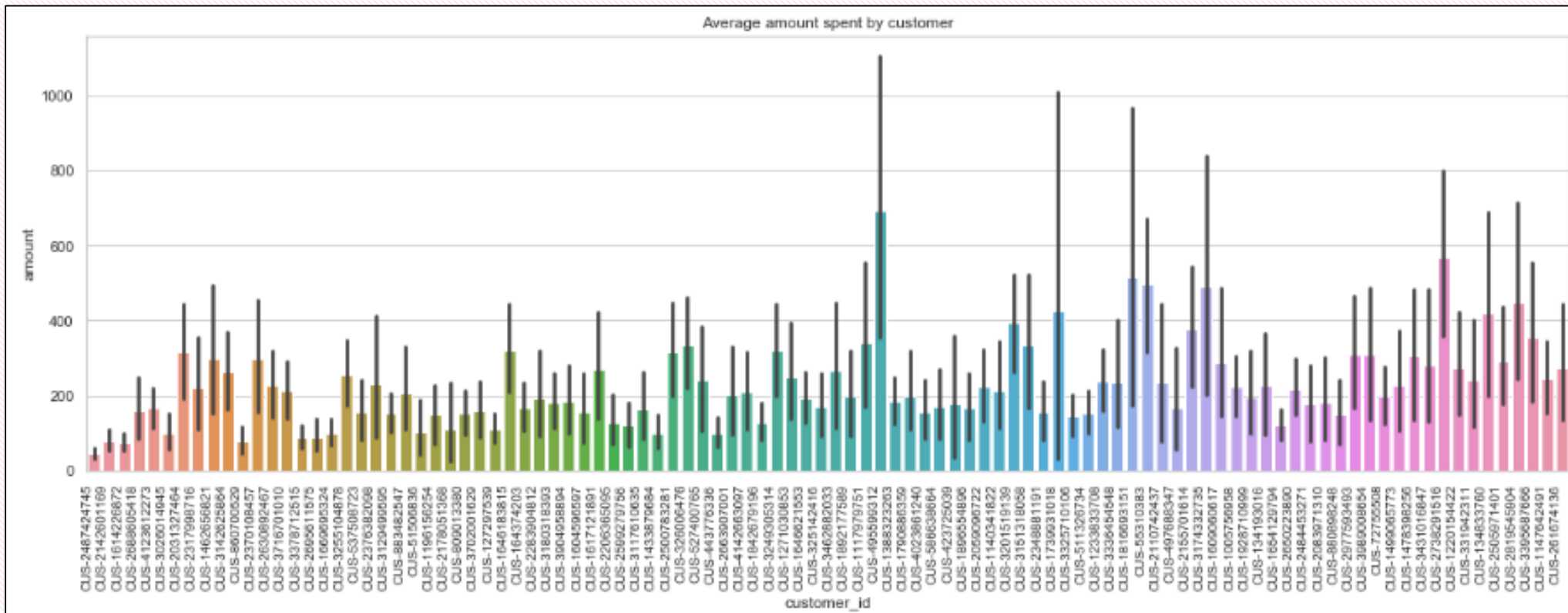
customer_id	
CUS-495599312	693.632903
CUS-2738291516	567.614500
CUS-1816693151	515.583846
CUS-55310383	497.669474
CUS-3174332735	489.220980
...	
CUS-3378712515	87.546000
CUS-2142601169	78.206106
CUS-860700529	77.681888
CUS-1614226872	74.465019
CUS-2487424745	45.348772

TOP 10 CUSTOMERS WITH MOST TRANSACTION

customer_id	first_name	
CUS-2487424745	Diana	578
CUS-2142601169	Michael	303
CUS-3026014945	Tonya	292
CUS-3378712515	Kimberly	260
CUS-1614226872	Rhonda	259
CUS-1196156254	Jessica	245
CUS-2695611575	Joseph	239
CUS-51506836	Richard	239
CUS-860700529	Jeffrey	233
CUS-1669695324	Tiffany	230

TOP 10 CUSTOMERS SPENDING THE MOST

customer_id	first_name	
CUS-2738291516	Kenneth	45409.16
CUS-3142625864	Ricky	42688.30
CUS-1816693151	Tim	40215.54
CUS-2155701614	Linda	37943.79
CUS-261674136	Kenneth	36786.13
CUS-883482547	Michelle	36639.41
CUS-4142663097	Sandra	36588.25
CUS-527400765	Natasha	36543.61
CUS-1196156254	Jessica	36050.72
CUS-2031327464	Michael	35832.97

TOP 10 RICHEST CUSTOMERS

| customer_id	first_name	
CUS-527400765	Natasha	29983336.71
CUS-2370108457	Patrick	17671502.33
CUS-3117610635	Christopher	8547297.74
CUS-2819545904	Brian	7011950.25
CUS-2663907001	Craig	6505599.14
CUS-1196156254	Jessica	5456746.27
CUS-1816693151	Tim	5102791.53
CUS-1609060617	Michael	4591549.46
CUS-2059096722	Michael	4226400.33
CUS-3462882033	Michael	3985131.11

CHALLENGE

Top 10 Areas with most balance

```
merchant_long_lat
151.21 -33.87     1549470.38
144.97 -37.81     1117726.11
144.96 -37.82     1022899.58
144.96 -37.81      866814.56
151.18 -33.8       687361.32
144.94 -37.6       634616.12
144.83 -37.78      619222.48
145.13 -37.78      607166.04
145.27 -37.48      592120.89
151.2  -33.87      591651.76
```

Top 10 Areas with most spending

```
merchant_long_lat
151.21 -33.87     12308.02
153.02 -27.48     11384.86
151.17 -33.94      9480.59
144.96 -37.82      6369.59
151.2  -33.87      4333.20
144.89 -37.77      3745.63
140.78 -37.83      3483.33
151.21 -33.86      3334.81
130.84 -12.44      3328.95
153.07 -27.53      2714.32
```
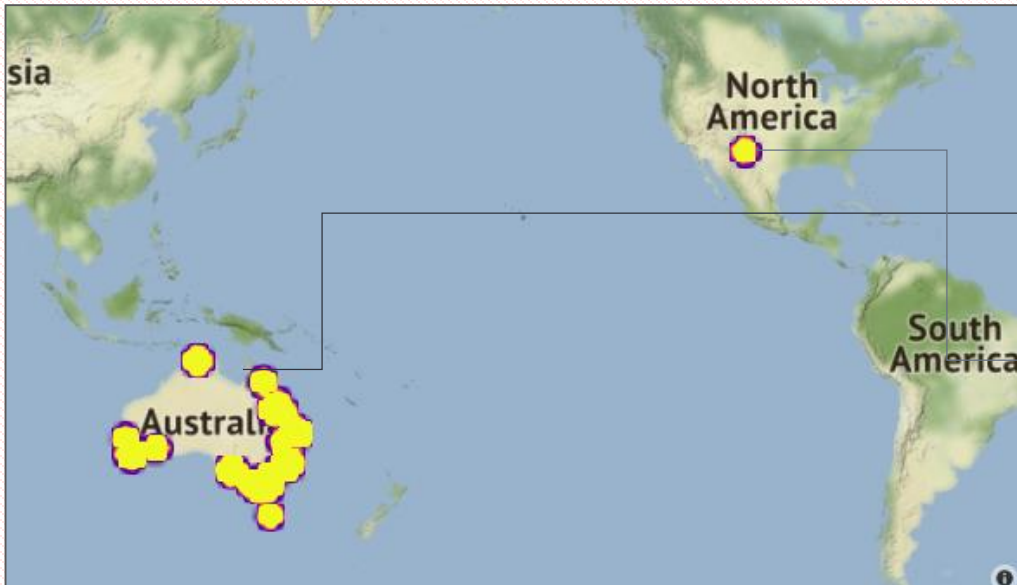

Most of the customer belongs to the coastline of Australia. However there were few customers who had a transaction coordinates of North America.
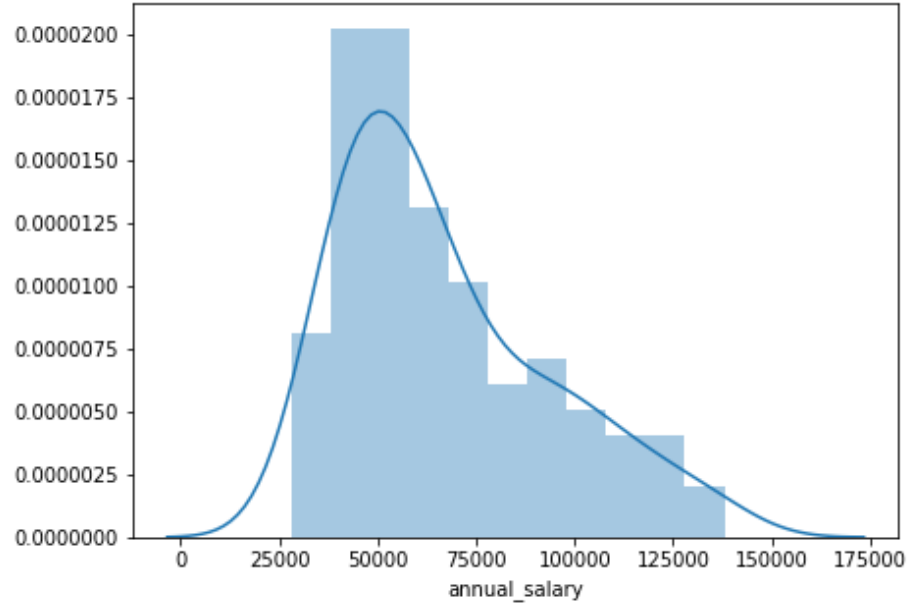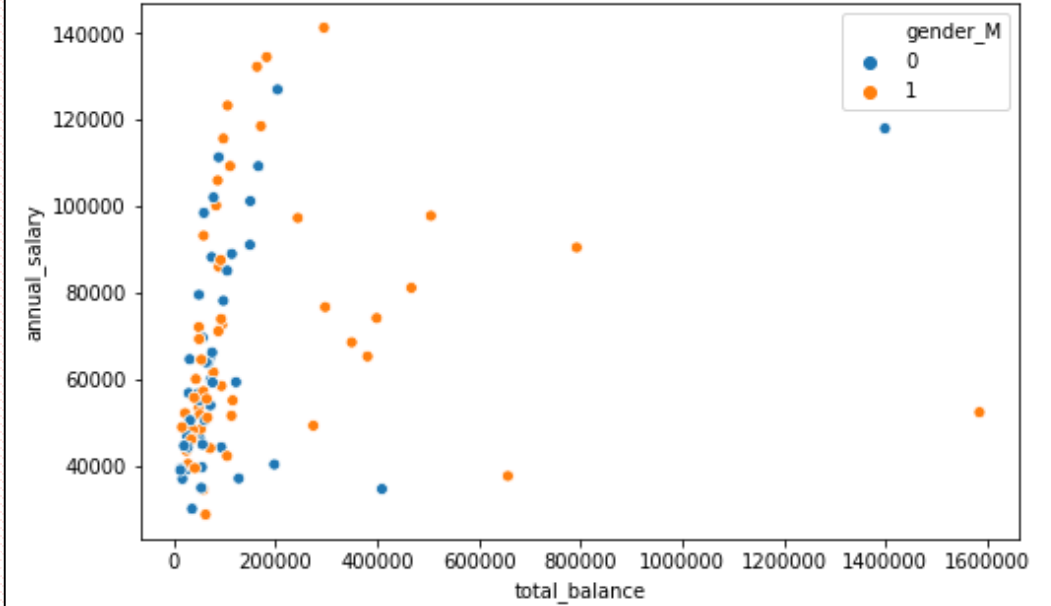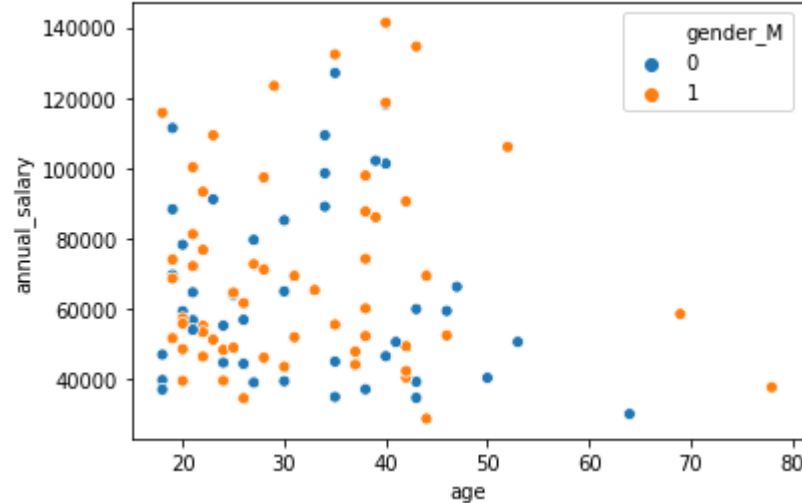
TASK -2

THE CORRELATION OF ANNUAL SALARY

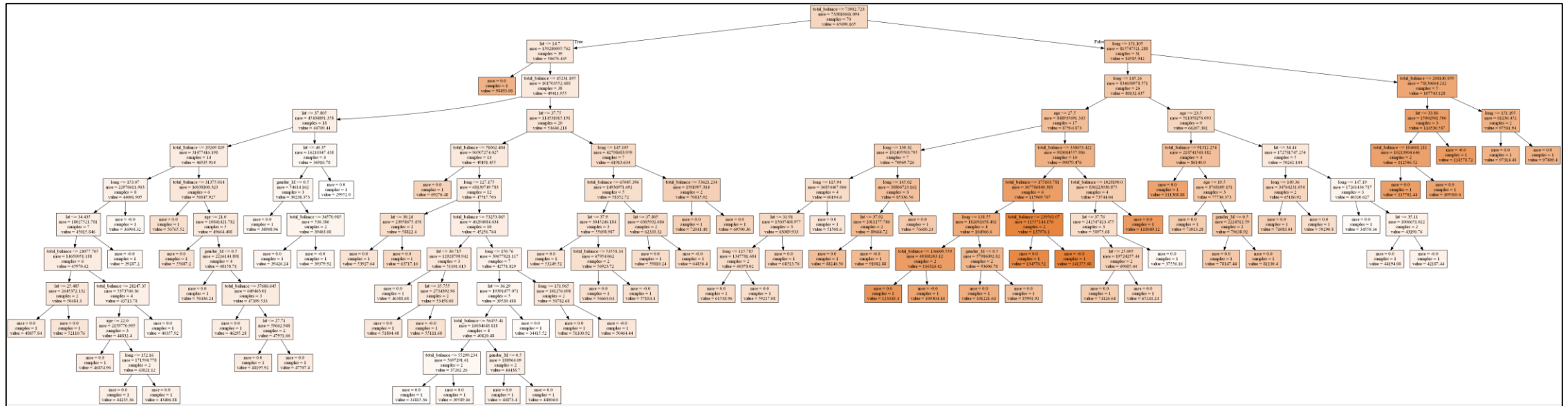# LINEAR REGRESSION MODEL

Annual_Salary VS Predicted (Training set)
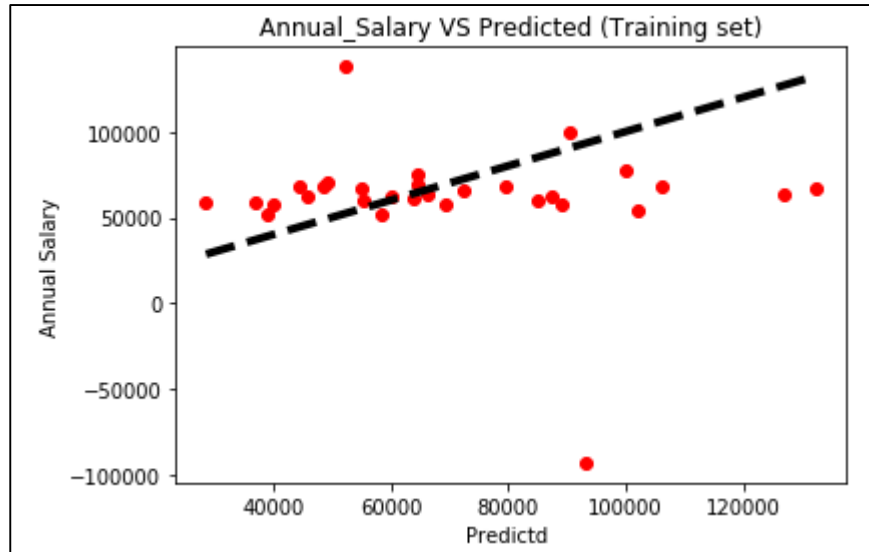
Linear Regression Model Conclusion:
- The linear regression model achieved an R-squared value of -0.42898854889679305.
- A Negative R2 means you are doing worse than the mean value, it indicates that the model explains very little of the variability of the annual salary.
- The model also achieved a Root Mean Squared Error (RMSE) of 45107.45412568924. It is the square root of the variance of the residuals. Residuals are a measure of how far from the regression line the data points are.
- In general, the lower the RMSE, the better the model is.
- Hence, this Linear Regression model has performed very poorly in predicting the annual salary of a customer.

CHALLENGE: DECISION TREE MODEL

Decision Tree Model Conclusion:
- The Decision Tree model achieved an R-squared value of 0.3937087218941056. As the R-squared value is closer to 0, it indicates that the model explains very little of the variability of the annual salary.
- On the other hand, the model achieved a Root Mean Squared Error (RMSE) of 38143.1869640057. This is much lessee than that of the linear regression model.
- Hence, although the Decision Tree model has performed better than the Linear Regression model in terms of its R-square value & RMSE. This means that the absolute fit of the model is much worse. Hence, I do not recommend using either models to segment the customers without improving the model. Some room for improvement include adding more relevant variables or using a different type of machine learning model.

THANK YOU

Contact me:

Jumana Nagaria
https://www.linkedin.com/in/jumana-nagaria/