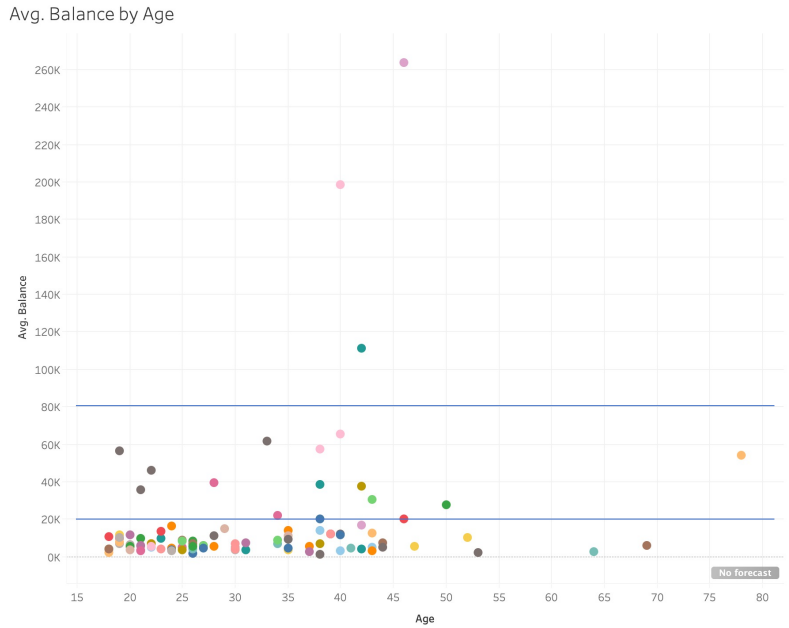
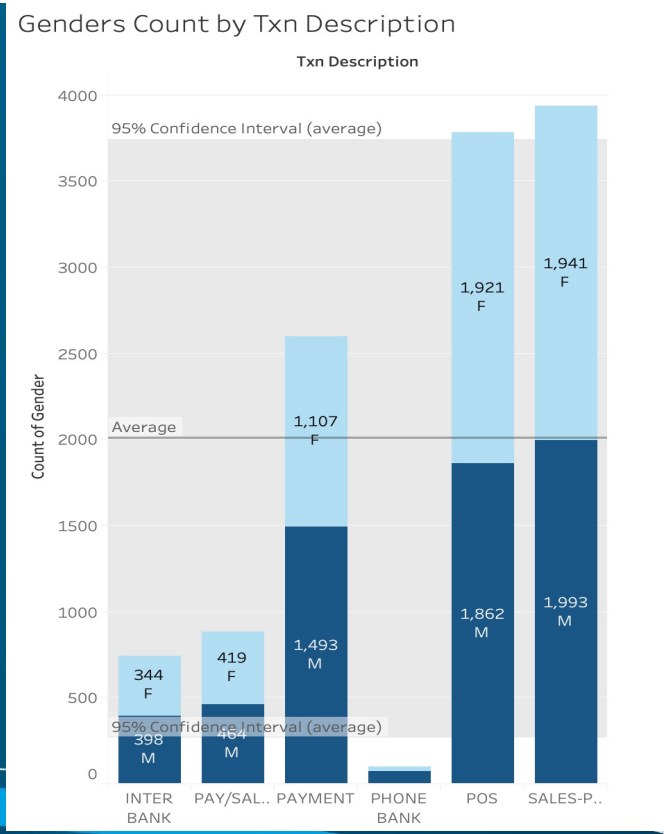
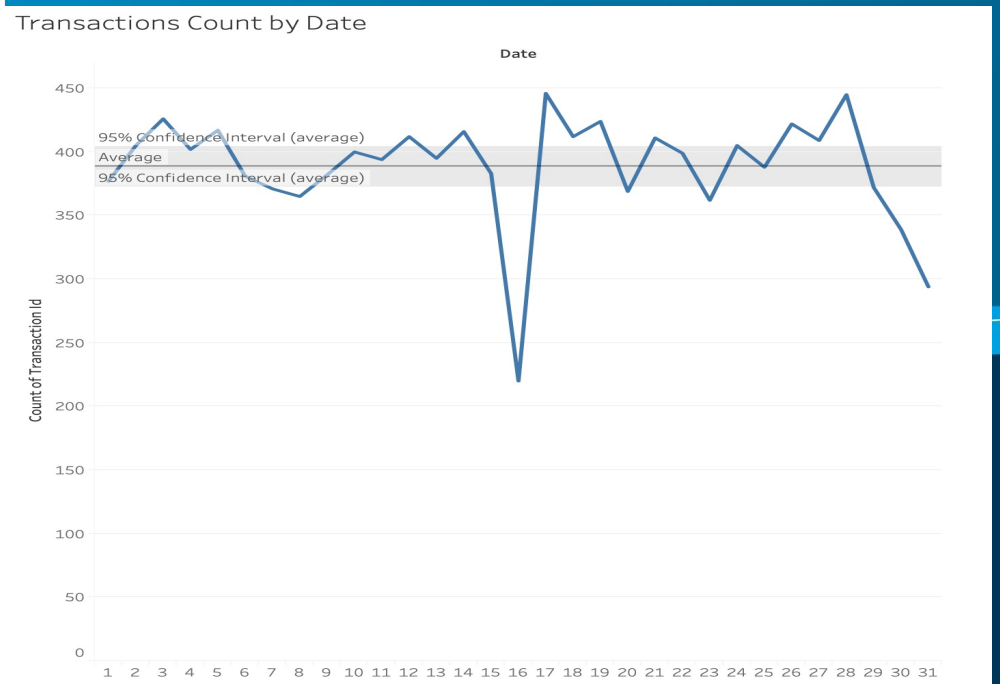
The first interesting point is the gender distribution for each transaction type. Overall, the phone bank recorded the smallest amount of transactions, but the point of sales recorded an enormous amount of transactions. Then we could tell that most of the customers preferred to make transactions at the point of sales, but we still need to do propaganda work on the phone bank.
Specifically, in each type of transaction, the female amount is slightly less than the male amount, except for the point of sales type. Hence we could conclude that actually, the male group made more transactions.



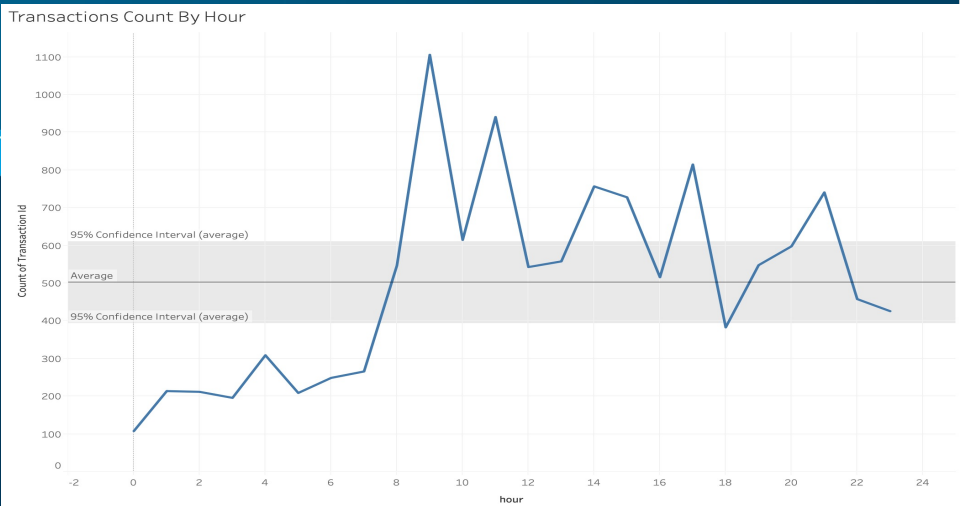
The second interesting point is the relationship between the average balance and age. And we could notice that most of the customers have an average balance of less than 20k. Then, some customers have a higher average balance of greater than 20K but less than 60K. Only three customers have an average balance of over 60K. And those three customers are ages between 40 and 50. We could treat those three points as outliers, but we need to make the decision later.



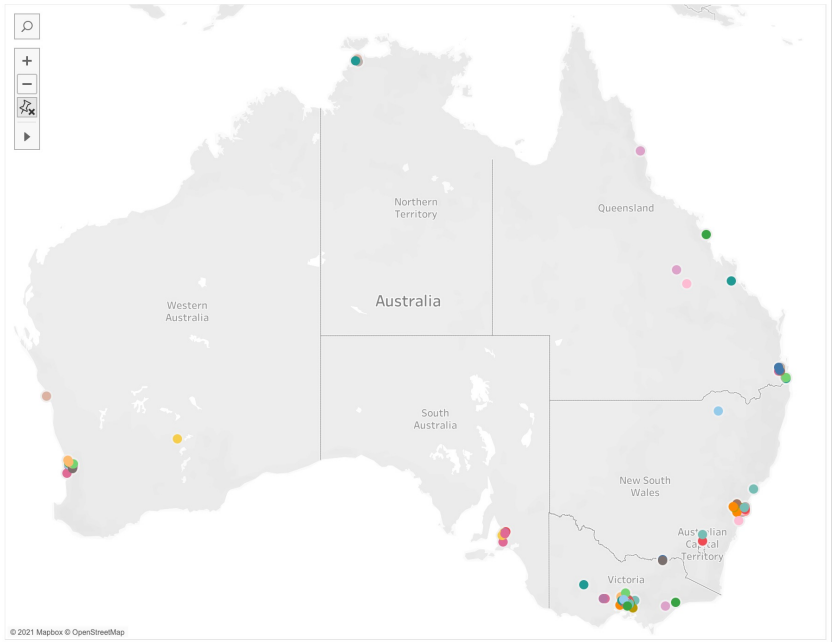
The third interesting point is when we show the transactions amount each day of the entire month. We can see that the daily amounts changed slightly, but the overall performance was constant. However, the amount decreased sharply on 16th, and then rapidly increased in the next day. We could figure out what happened on that specific day. Then, at the end of the month, the daily amounts started to decrease, customers probably prepare to save money for the next month. And the average amount of transactions is roughly 380.



The fourth exciting point is when we count the total transactions amounts by hours. From the plot, we could notice four peak values that we could roughly get five clusters. The first cluster starts from midnight to 8 am, and customers are sleeping and start their morning routine. The second cluster includes two peak values, which shows the transactions amounts increased rapidly, and customers became active. The third cluster is the duration between noon to 4 pm, and the transactions amounts decreased a lot. Then, a third peak value appeared between 4 pm and 6 pm, which means the transactions increased in this duration. The last cluster contains the last small peak value, which means some customers enjoyed night shopping. And the average transactions per day is 500.



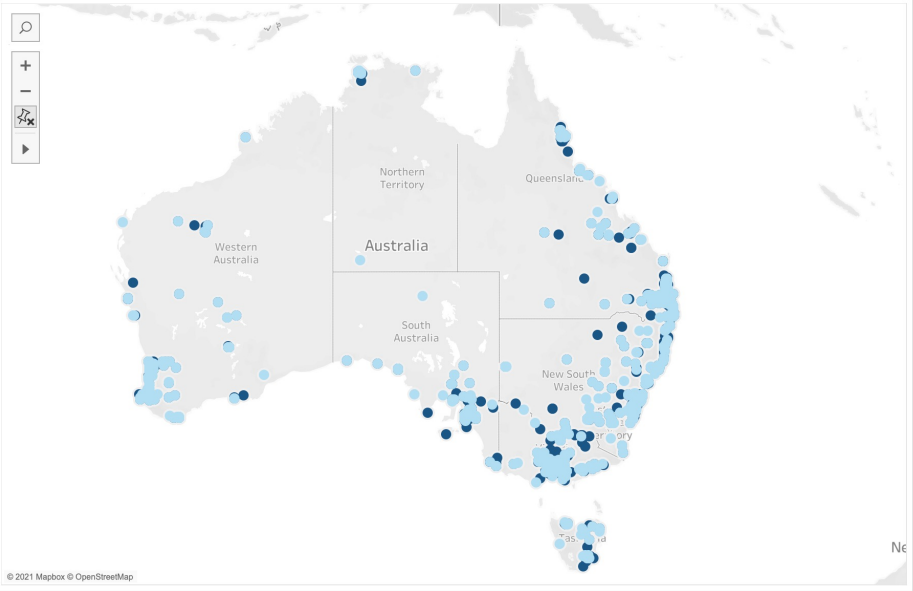
Customers Location Map



Customer Id
- CUS-51506836
- CUS-55310383
- CUS-72755508
- CUS-127297539
- CUS-134193016
- CUS-134833760
- CUS-164374203
- CUS-261674136
- CUS-325142416
- CUS-326006476
- CUS-331942311
- CUS-423725039
- CUS-443776336
- CUS-495599312
- CUS-497688347
- CUS-511326734
- CUS-527400765
- CUS-537508723
- CUS-586638664
- CUS-809013380
- CUS-860700529
- CUS-880898248
- CUS-883482547
- CUS-1005756958
- CUS-1117979751
- CUS-1140341822
- CUS-1147642491
- CUS-1196156254
- CUS-1220154422
- CUS-1233833708
- CUS-1271030853
- CUS-1388323263
- CUS-1433879684
- CUS-1462656821
- CUS-1478398256

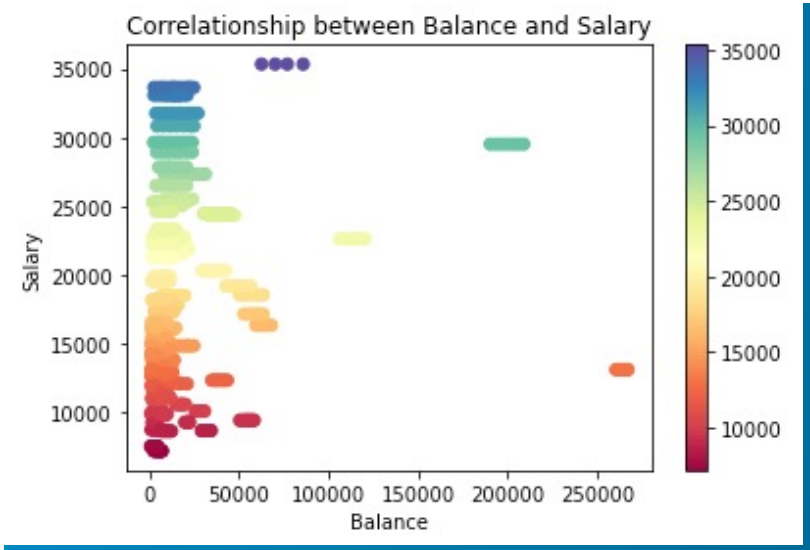
The fifth exciting point is when we plot customers' locations on the map. We plot the customer ID on the map, which is unique for each customer. And we could see that most of the customers located in Melbourne and Sydney. Then, some customers located in Perth and Brisbane. Only severe customers in Darwin and Adelaide. The distribution of customers follows the population of each state. Since Victoria and NSW have many populations, more customers will be in those two states.

Merchant Location By Transaction Description

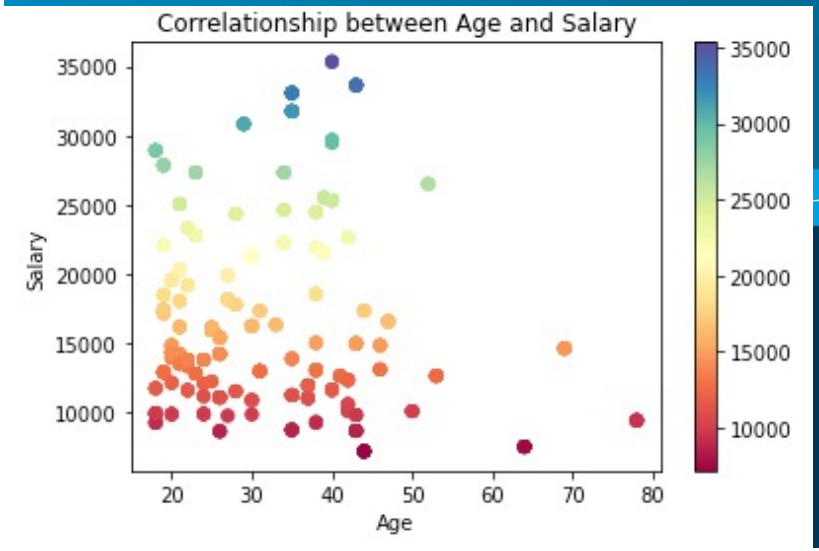


Txn Description
- POS
- SALES-POS

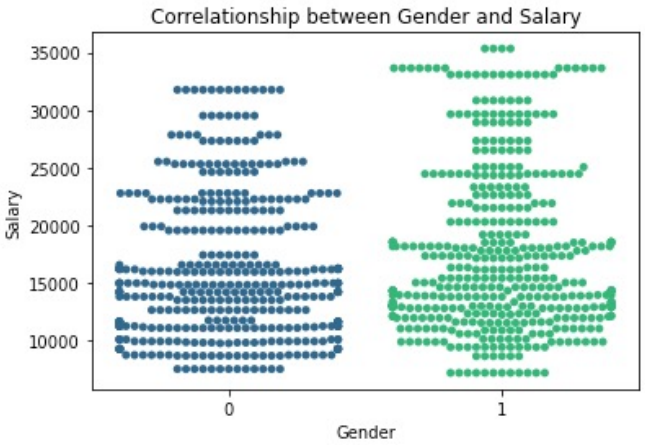
The last interesting point is the plot of merchant location by transaction description. We could notice that there are vast amounts of merchant locations around the whole of Australia. But only POS and Sales-POS will be recorded in each merchant location. And POS locations are far more than Sales-POS locations.



The first plot indicates the correlations between balance and salary for each customer. Balance kindly relates to customers' purchasing behavior. And generally, a customer with a high salary will have a high balance. However, we could notice from the plot that the correlation between balance and salary is rarely low. It represents that most customers spent a lot but saved less.



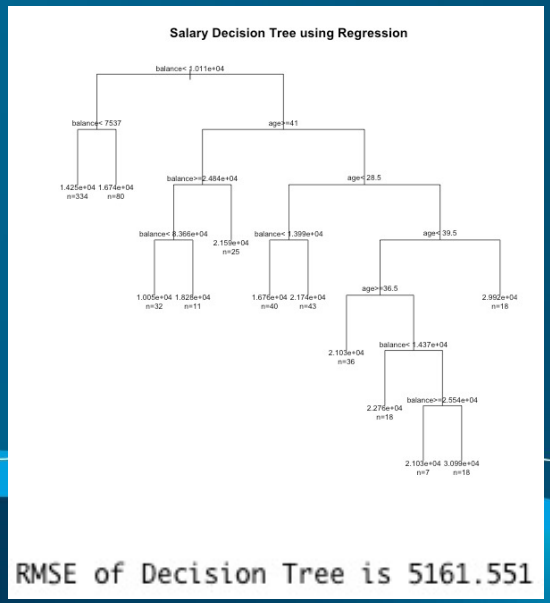
The second plot indicates the correlation between age and salary. And we could see that there are no apparent relationships between those two variables. The age of customers whose salary is over 30K are all under 50. And along with age increasing, the salary would decrease, and the salary gap between customers is not significant. However, when the age is lower than 50, the salary levels vary from 10K to 35K. And this symptom may be determined by working skill levels, educations, and corporations' sizes.



The third plot indicates the correlation between gender and salary. One represents Male, and Zero represents Female. From the plot, Male customers obviously could get a high salary. And the highest salary for male customers is 35K, but the highest salary for female customers is roughly 32K. Overall, female customers receive a lower salary compared to Male customers.

```
Linear Regression

883 samples
  3 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 882, 882, 882, 882, 882, 882, ...
Resampling results:

  RMSE      Rsquared    MAE
  6572.592  0.04112573  5196.686

Tuning parameter 'intercept' was held constant at a value of TRUE
```

We build a simple regression model using indicators: balance, age, and gender to predict salary. From the summary table, we could see that this linear model does not perform well. The R-squared value is relatively low, with only roughly 4.11% data explained by the model. And the RMSE value is very high. Therefore, the simple model could not be used to segment following customers.



RMSE of Decision Tree is 5161.551

We used cross-validation to separate the train data and test data to build a decision tree model for regression. And the performance of this model improved compared to the simple model. The RMSE value is 5161.56. Hence, this model performs better.