Data Science Report

Author**:**

* Hendrik Schmidt (hendrikmaximiliangerhard.schmidt@student.uts.edu.au)

# Toolset:

* R/ R-Studio: dplyr, ggplot2, readr, tidyr, ISLR, hydroGOF, WriteXLS

# Business Understanding

The objective of this project was to accurately predict the monthly total transaction amount for the next month by a given industry location combination.

# Data Understanding

The data was provided in the form of CSV file through the financial services company.

It comprised 94248 observations of 5 variables and covert a period from January 2013 till November 2016. This period led to the conclusion that the prediction aims the December of 2016.

The first part was to convert the variables of the dataset to a manageable format and figure out if all variables (columns) were necessary for the prediction.

The second part dealt with the identification of the subsets of the data on which the modelling and prediction would be based on.

# Data Preparation

The main tasks in this phase were the transformation of the data, the visualization of the raw data by the representation of variation and diagrams and finally the examination and adjustment of a linear regression.

The vectors industry and location were encoded as a factor. Moreover, an aggregated data set with a mean of the monthly transaction amount was created and the customer id was dropped, because of the insignificance for the prediction.

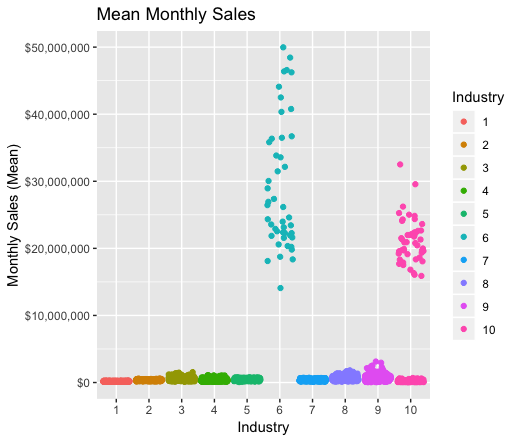
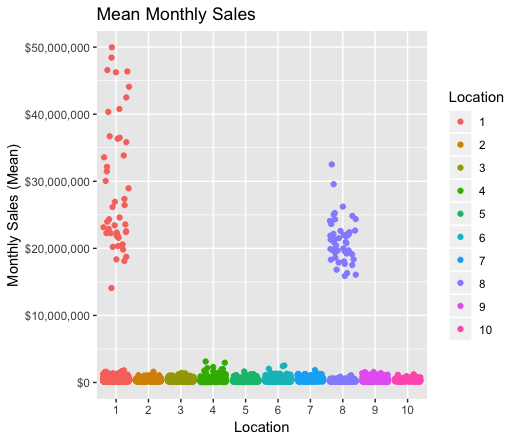
During the exploratory data analysis (EDA) outliners were detected. Especially the mean of the monthly transaction amounts of industry 6 and industry 10 were widely spread in comparison to the other observations. The total values were significantly higher as well (see Graph 1). The same applied for location 1 and location 8 (see Graph 2).

The combination of industry and location with the highest deviation and significantly higher total values were found within industry 10 and location 8 and industry 6 and location 1.

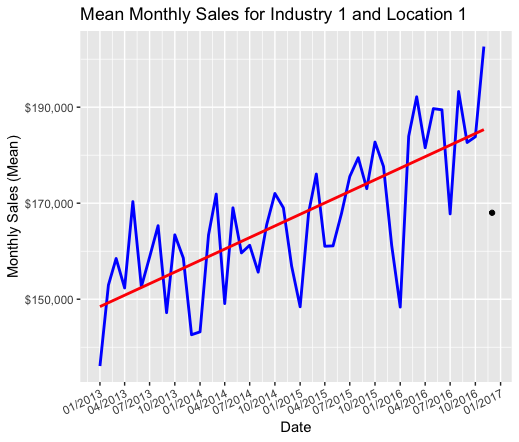
As an example for seasonality within the data, the subset industry 1 and location 1 showed how at the end of every year the monthly total transaction amount spiked downwards (see Graph 3).

To account this seasonality, two new variables were generated, one for the month and one for the year. The variable which accounts for the months was set up as a factor, which means that

January = 1, February = 2 etc. The year variable was assigned an integer value.



Graph 1: Graph 2:



Graph 3:

# Modelling

The specific modelling technique is based on a linear regression. Moreover, since multiple variables were needed to enhance the regression, a multiple linear regression was used.

To be able to use a multiple linear regression for the prediction and test its performance, a train and test set was needed.

For the training set, the first 36 months of data were used. The testing set included the remaining eleven months. The regression was fitted based on the two variables.

These variables capture both the seasonality and the trend of the monthly amount for the given industry/location combination, as seen as an example for industry 1 and location 1 in graph 3.

A cubic polynomial was applied to the year variable to get a minimal increase in the R- squared estimator, which describes the overall fit of the model from a scale of 0 to 1.

By applying the regression based on the two features, the overall performance (based on the test set) with an adjusted R-squared of 0.7458 and an RSME of 11298.02, showed a reasonable fit for the model.

The RSME measures the differences between the predicted model of the training set to the test set. The target is to reduce the RSME which would mean a reduction in the difference of the residuals (error terms).

Furthermore, the predicted total transaction amount for December 2016 fitted the subset of industry 1 and location 1 with $168,016 sufficing (see graph 3, black point). In addition, 95% of the prediction interval of the transaction amount for the December 2016 is between $153,866 and $182165.3 for industry 1 and location 1.

# Evaluation

The initial data mining has revealed interesting features about the overall performance of certain industry location combination. However, the task is to accurately predict the monthly total transaction amount for December 2016 over all different industry location combination.

The model performed very well for most of the entire data. For example, the combination of industry 8 and location 8 performed with an adjusted R-squared of 0.934 and high significance towards the variables, very well.

On the downside, the model performed poorly for the combinations of industry 3, location 8 & industry 5, location 3. Both adjusted R squared estimator are negative which comes from the complete insignificance of the feature variables. Other selected features variables would maybe perform for these subsets differently.

Given a more diverse data set in terms of more innate variables from the outset, the model could potentially perform better.

In conclusion, the model has the capability to deliver an accurate result for the monthly amount for December 2016 (see Appendix).

# Deployment

The model could be deployed as a prediction tool to support and help develop future predictive models within the company.

The model’s results need to be checked by a data scientist to ensure the predictive accuracy and that it can be used for upcoming predictions.

# Appendix

