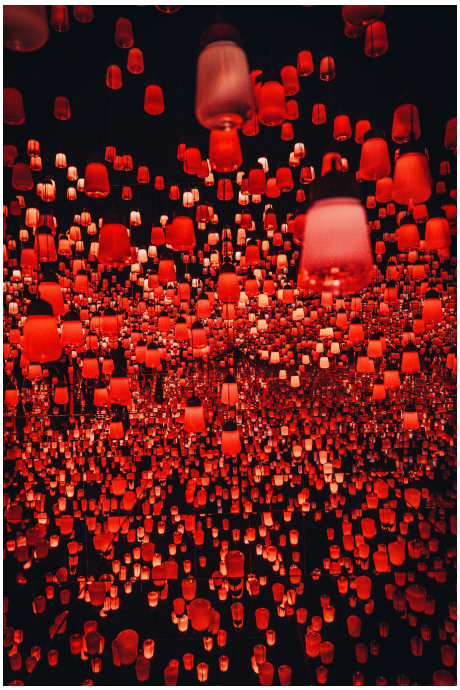

ETC513 Assignment 4: Practical Driving Tests in Queensland

Shaohu Chen
Master of Business Analytics(In Progress)

Qian Duan
Master of Business Analytics(In Progress)

Tina Tsou
Master of Business Analytics(In Progress)

Report for
Queensland Government

3 June 2021

Practical Driving Examination Board

(03) 9905 2478
questions@company.com

ABN: 12 377 614 630

Introduction

Data Description:

This data set is Practical driving examination results for customers which is provided by local government authority (LGA) of Queensland. It records the license class, booking type, examination results and driver age group during 2005 to 2019.

Research aims:

We divided into three parts, the first part focuses on the annual pass rate of different local government authority.

The second part mainly aims to compare the age group with different license.

The third part calculates the correlation between the examination results and booking type.

Using R Core Team (2013), we ran analysis to explore our research goals. This analysis uses R packages Wickham et al. (2019), Wickham and Hester (2020), Zhu (2021), Xie (2021), Wickham (2016), Kaplan (2020), Wickham and Seidel (2020), **stargazer**, Robinson, Hayes, and Couch (2021), and Robin et al. (2011).

Part 1

The first part focuses on the annual pass rate of different local government authority.

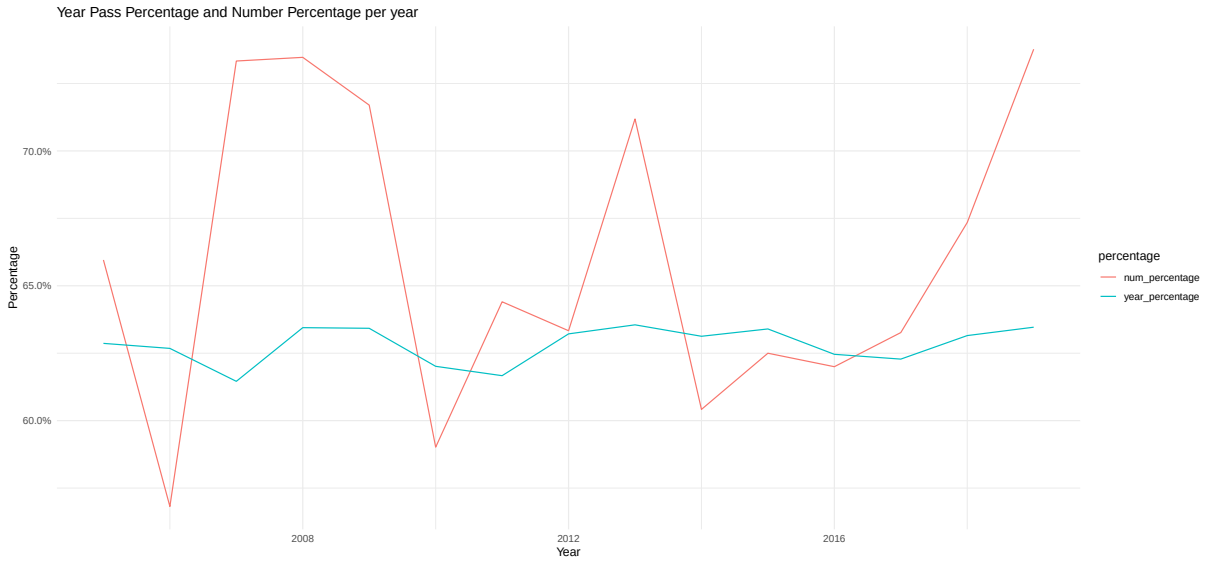

Figure 1: *Year Pass Percentage(year_percentage) and Percentage of Local Government Authority annual passing rates exceeding the total annual passing rate(num_percentage)*

In the figure 1, as for the annual pass rate, it does not fluctuate greatly, and basically remained at 62.5% but the percentage of the number exceeding the annual pass rate fluctuates greatly, which may be due to missing data in some regions, however, from the data point of view, it has been in an upward phase in recent years.

Table 1: *Number of time getteing the highest pass rate per year*

Local Government Authority	count
BLACKALL-TAMBO REGIONAL COUNCIL	5
BALONNE SHIRE COUNCIL	4
HOPE VALE ABORIGINAL SHIRE COUNCIL	4
MOBILE SERVICES	4
BARCALDINE REGIONAL COUNCIL	3
BURKE SHIRE COUNCIL	3

Table 2: *Number of time getteing the lowest pass rate per year*

Local Government Authority	count
MAREEBA SHIRE COUNCIL	2
NAPRANUM ABORIGINAL SHIRE COUNCIL	2
REDLAND CITY COUNCIL	2
BURDEKIN SHIRE COUNCIL	1
HOPE VALE ABORIGINAL SHIRE COUNCIL	1
KOWANYAMA ABORIGINAL SHIRE CONUNCIL	1

In the table 1, BLACKALL-TAMBO REGIONAL COUNCIL has the most number of first (5 times). In the table 2, MAREEBA SHIRE COUNCIL, NAPRANUM ABORIGINAL SHIRE COUNCIL, REDLAND CITY COUNCIL have won the last 2 times.

According to the figure 2, the annual pass rate of BLACKALL-TAMBO REGIONAL COUNCIL has been on the rise after 2007, even reaching 100%, while the annual pass rate of MAREEBA SHIRE COUNCIL is in a downward state, and the annual pass rate of REDLAND CITY COUNCIL basically fluctuates at 55%. Especially, MAREEBA SHIRE COUNCIL and REDLAND CITY COUNCIL have been lower than the annual pass rate since 2015, which the gold line is year pass rate.

In the figure 3, the number of pass for three government authorities basically has little fluctuation. Interestingly, the ranking of the number of passes and the ranking of the pass rate are completely opposite.

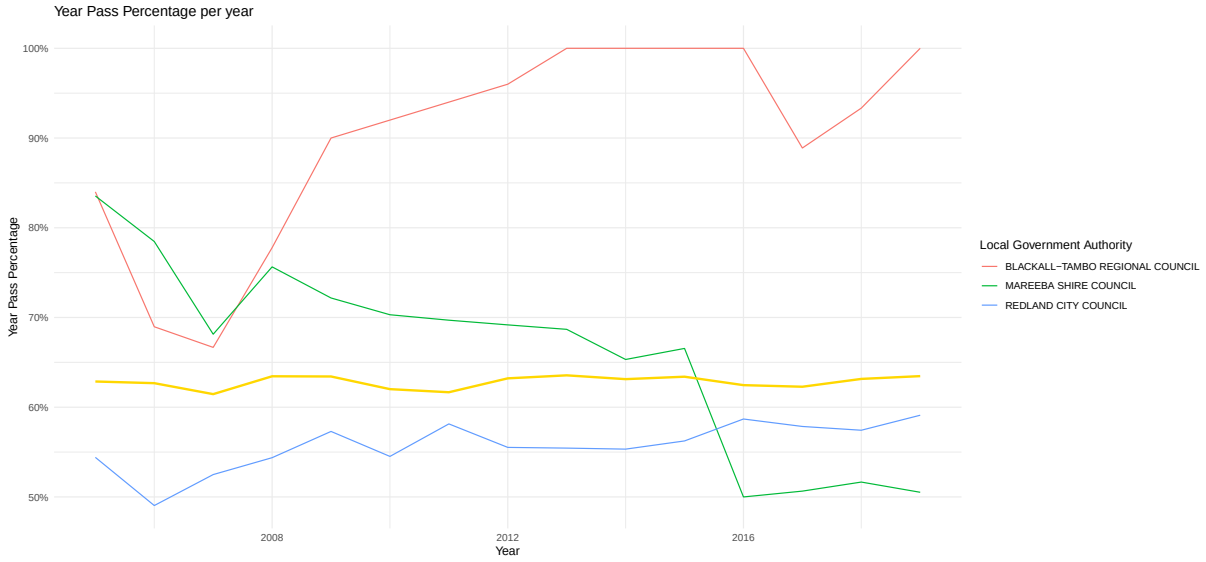

Figure 2: *Year Pass Percentage in BLACKALL-TAMBO REGIONAL COUNCIL, BLACKALL-TAMBO REGIONAL COUNCIL, and REDLAND CITY COUNCIL*

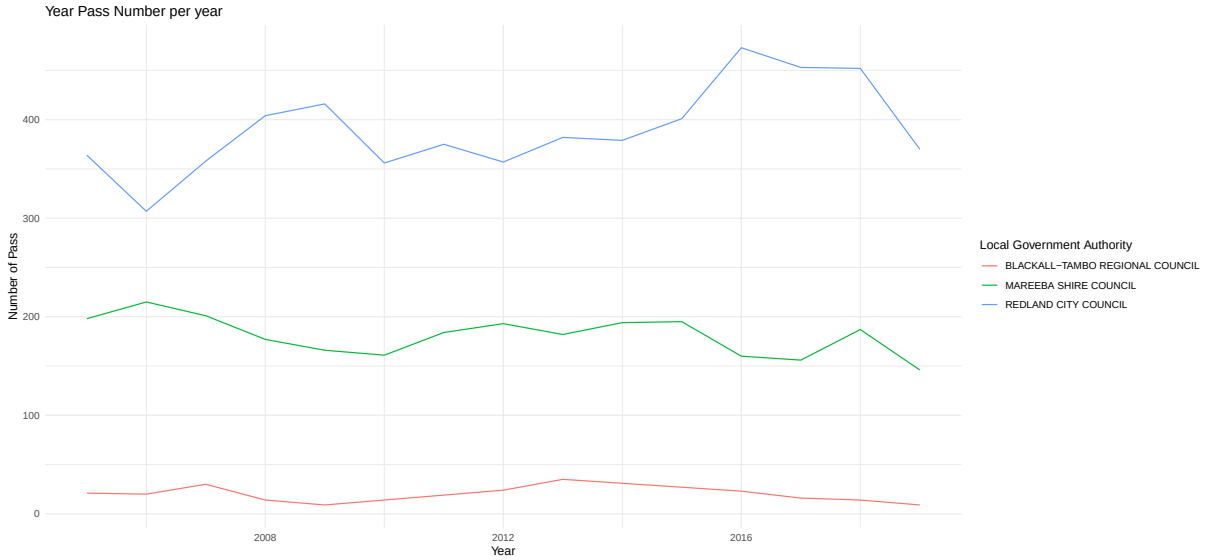

Figure 3: *Year Pass Number in BLACKALL-TAMBO REGIONAL COUNCIL, MAREEBA SHIRE COUNCIL, and REDLAND CITY COUNCIL*

Therefore, the annual pass rate has not changed much, the percentage of the number exceeding the annual pass rate fluctuates greatly. Areas with sparsely populated areas may have fewer people participating, resulting in a higher overall pass rate than areas with densely populated areas.

Part 2

The driving examination pass rate of Queensland is 63% .

Table 3: *The pass rate of each product type*

Product Type Name	pass_rate
CLASS CA - CAR (AUTOMATIC)	53%
CLASS C - CAR (MANUAL)	56%
CLASS HR - HEAVY RIGID VEHICLE	70%
CLASS MR - MEDIUM RIGID VEHICLE	77%
CLASS RE - MOTORCYCLE (UP TO 250CC	77%
CLASS HC - HEAVY COMBINATION VEHICLE	79%
CLASS LR - LIGHT RIGID VEHICLE	83%
CLASS R - MOTORCYCLE (OVER 250CC)	86%

- The table 3 is the pass rate of different licenses. Above all, the automatic car has the lowest passing rate with 53%, Since the car has the largest amount of popularity to meet people daily command, so there are more people to join the test of cars.
- While the motorcycle over 250cc has the largest pass rate with 86%. The motorcycle is much more professional, and it required people got the license up to 250cc who can take part in the test. Therefore, these people are professional so got a higher rate.

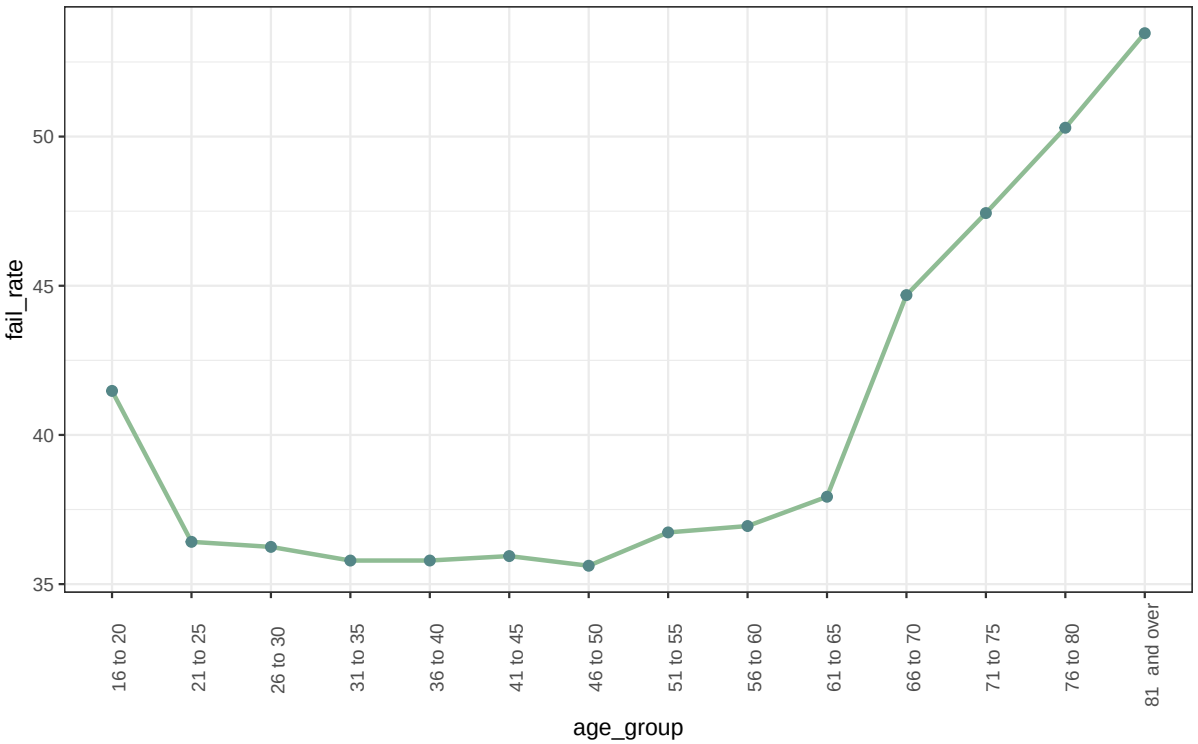

Figure 4: *Queensland driving test fail rates by age*

Figure 4, describes the failing rate of different age groups. It can be seen that the fail rate is increasing with the age grows older. Because people is 81 years old and over has the highest fail rate, which

means it is hard for people to pass the driving license after 61 years old. However, there is one interesting point for young people with high rate at 41%. Basically, the people in this range takes the highest number of examinations. While in the original data set, some young people around these ages have 200- or 300-times test, but still failed. Australia government has more restrictions on young people driving license, so the rate is high.

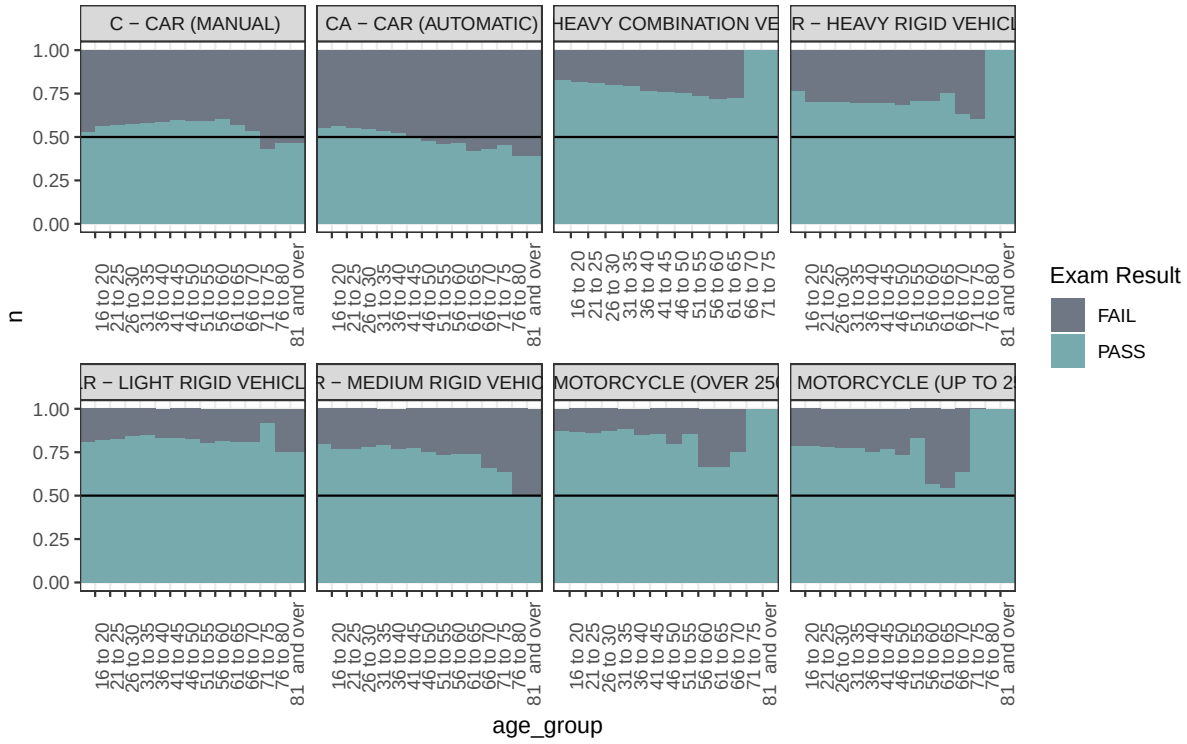

Figure 5: *Compare the fail and pass in different license*

In Figure 5, for most drive license, the number of pass all exceeds 50% of all age, except the automatic car license. People who fails at an older age. The number of fails is much more than pass. Another interesting point is that in motorcycle and heavy vehicle, people over 70 years old get 100% pass rate. This is mainly because there are only one or two people join the test and he pass. So, the pass rate is 100%. This doesn't mean all old people can get the license for one time.

Part 3

In the following section, we will be analyzing the relationship between *Booking Type* and *Exam Result*.

The frequency plot, Figure 6, between *Booking Type* and *Exam Result* shows that the percentages of people who passed the exam are similar for both driving school and private.

Figure 6: *Frequency Plot between Booking Type and Exam Result*

Since the response variable and predictor variable are categorical variables, they will have to be converted into dummy variables(0 & 1). Then, following regression analysis Alice (2018), I ran a logistic regression to analyze their relationship. The following is the formula for the regression *logmodel*:

$Y\ B(p), log(\frac{p}{1-p}) = \beta_0 + \beta_1 X + \epsilon$

- β_0 is the intercept.
- β_1 is the coefficient of *Booking Type_Private*
- X is *Booking Type_Private* taking values 0 or 1

Table 4 shows the regression summary. *Booking Type_Private* has p-value close to 0 which means it is statistically significant. Due to the variable being a dummy variable relative to booking type driving school, the coefficient indicates that *Booking Type_Private* affects the passing of an exam negatively compared to *Booking Type_Driving School*. Private booking reduces the log odds by 0.061.

Table 4: *Regression Result for logmodel*

	Dependent Variable:
term	'Exam Result_PASS'
(Intercept)	0.533407196218835 ***
'Booking Type_Private'	-0.0179325664251033 *
nobs	337084
AIC	444926.751217685
logLik	-222461.375608843

Note:
makecell[l]p-value: '+' 0.05 - 0.1; '*' 0.01 - 0.05; '**' 0.001 - 0.01; '***' 0 - 0.001

Table 5: *Anova for logmodel*

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	337083	444929.1	NA
'Booking Type_Private'	1	6.315064	337082	444922.8	0.0119716

ANOVA test, Table 5, on the *logmodel* analyzes the table of deviance which shows how well the x variable is doing in comparison to the null model. Here we can see that the drop in deviance is quite small despite having low p-value.

Next, we test the fit of the model by looking at the receiver operating characteristic (ROC) curve.

Figure 7 shows the ROC curve of the *logmodel*. It is basically a 45 degree diagonal line which indicates the model has no discrimination ability.

Area under the curve "...gives the probability that the model correctly ranks such pairs of observations" Bartlett (2014). The area under the curve for this model is 0.5022363. In conclusion, the predictor just makes random guesses.

We try to improve the model by adding more variables to the function:

$Y \sim B(p)$, $log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

- X_2 is *Number of Examinations* taken by each examinee.

Table 6 shows the two regression summary side-by-side. The regression with *Number of Examinations* has AIC of 333404. It is slightly lower than the AIC of the previous regression which was 333483. Thus, in comparison, having this one extra variable improved the function significantly (statistically).

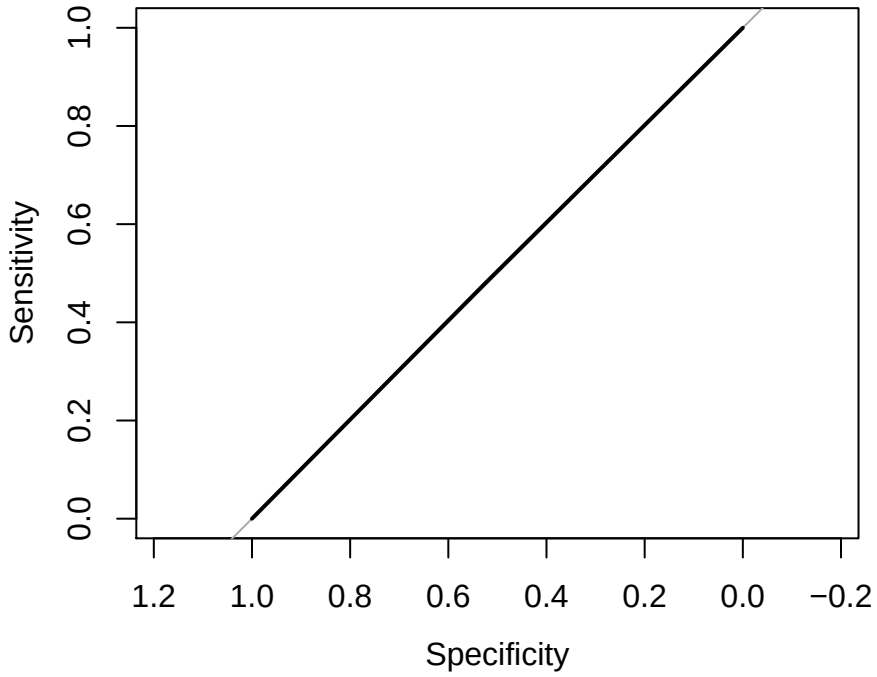

Figure 7: *ROC Cuve of logmodel1*

Table 6: *Regression Result Comparison*

	Dependent Variable:	
term	'Exam Result_PASS' (logmodel2)	'Exam Result_PASS'(logmodel)
(Intercept)	0.522573474570028 ***	0.533407196218835 ***
'Booking Type_Private'	-0.0157422192951367 *	-0.0179325664251033 *
'Number of Examinations'	0.00169271410738006 ***	
nobs	337084	337084
AIC	444885.718925042	444926.751217685
logLik	-222439.859462521	222461.375608843

Note:
makecell[l]p-value: '+' 0.05 - 0.1; '*' 0.01 - 0.05; '**' 0.001 - 0.01; '***' 0 - 0.001

However, a simple calculation of area under ROC curve for logmodel2,0.4859687,indicates that the model is even worse than the first.

Conclusion

In conclusion, we've shown that higher pass rate in certain districts is not always an absolute reflection on whether the district has better driving program. Rather, it is an outcome of locations with lower

examinees in general. For locations with more examinees, there would be more variations in their outcome thus more fails.

Next, we shown that automatic cars have the lowest pass rate overall, and that motorcycle (over 250cc) has the highest pass rate. Older people (66 and above) also tend to fail their vehicle tests more. But ultimately pass rate for each vehicle type and majority of the age group is over 50%.

Last but not least, although, there is statistical relationship between the booking type and the exam outcome, the affect is pretty small. Furthermore, the current variables are inadequate in creating a good model to predict the outcome.

This is also a shortcoming with the data we currently have. Because it contains very limited variables, it is hard to create a better fit model that can predict the outcome accurately.

References

Alice, M (2018). *How to Perform a Logistic Regression in R*. https://datascienceplus.com/perform-logistic-regression-in-r/.

Bartlett, J (2014). *Area under the ROC curve – assessing discrimination in logistic regression*. https://thestatsgeek.com/2014/05/05/area-under-the-roc-curve-assessing-discrimination-in-logistic-regression/.

Kaplan, J (2020). *fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables*. R package version 1.6.3. https://CRAN.R-project.org/package=fastDummies.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/.

Robin, X, N Turck, A Hainard, N Tiberti, F Lisacek, JC Sanchez, and M Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77.

Robinson, D, A Hayes, and S Couch (2021). *broom: Convert Statistical Objects into Tidy Tibbles*. R package version 0.7.6. https://CRAN.R-project.org/package=broom.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H, M Averick, J Bryan, W Chang, LD McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, D Vaughan, C Wilke, K Woo, and H Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686.

Wickham, H and J Hester (2020). *readr: Read Rectangular Text Data*. R package version 1.4.0. https://CRAN.R-project.org/package=readr.

Wickham, H and D Seidel (2020). *scales: Scale Functions for Visualization*. R package version 1.1.1. https://CRAN.R-project.org/package=scales.

Xie, Y (2021). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.22. https://github.com/rstudio/bookdown.

Zhu, H (2021). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra.