ST447 Data Analysis and Statistical Methods: Individual Project

Candidate Number: 85895

TABLE OF CONTENTS

INTRODUCTION

XYZ is a student from LSE who has been learning driving for some time, and is thinking of taking the practical car test in UK. XYZ's skill is about average.

XYZ has two options when it comes to choosing the location of the test center.

- Take the practical test at the nearest test centre to his/her home;
- Take it at the nearest test centre to the LSE.

This report helps to statistically determine which location is best suited for XYZ to maximize his/her chance of passing the car test.

PROFILE OF XYZ

XYZ's profile was pseudo-randomly generated using the function XYZprofile in R where the input argument is the numerical value of my 9-digit LSE Student ID. Using this function, XYZ has the following characteristics:

- Age: 22
- Gender: Male
- Home address: Rochdale (Manchester)

DATA SOURCE

The dataset used for this analysis was DVSA1203, which contained information on car pass rates by age (17 to 25 year olds), gender, year (2007-2015) and test centre location. The dataset is available at https://www.gov.uk/government/statistical-data-sets/car-driving- test-data-by-test-centre.

SCOPE OF REPORT

This report aims to answer the following three questions using statistics and data analysis:

1. What is XYZ's expected passing rate at the nearest test centre to his home?
2. What is XYZ's expected passing rate at the nearest test centre to the LSE?
3. Of these two locations, where should XYZ take the test?

ANALYSIS

PRE-PROCESSING AND CLEANING DATA

The data was analyzed using the programming language R. The dataset was of '*.ods' format which required some preprocessing before the data could be used in R. The data was first converted into a Microsoft Excel file, and which was then pre-processed using R.

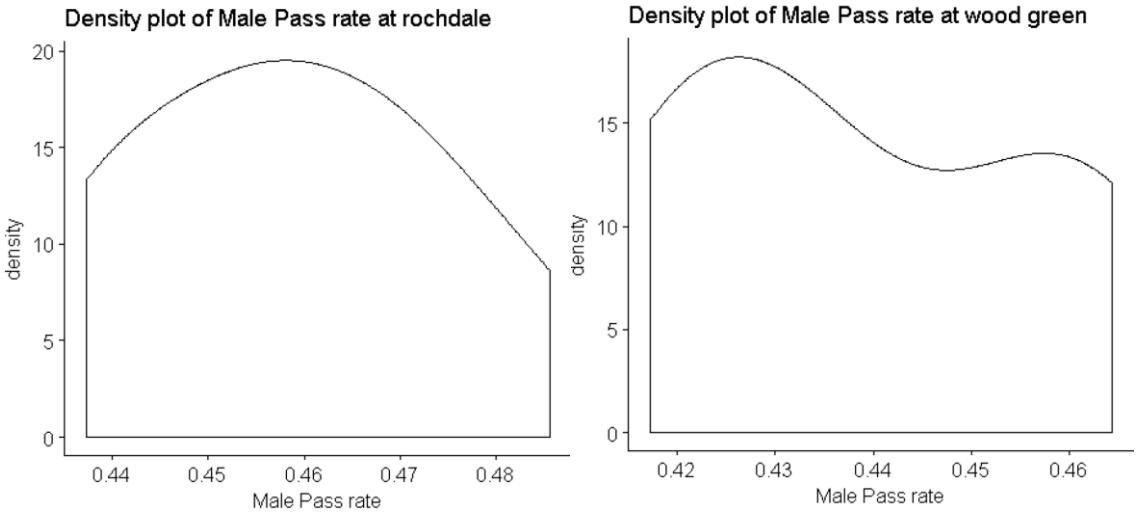
	Location	Age	Male_Conducted	Male_Passes	Male_Pass_rate	Female_Conducted	Female_Passes	Female_Pass_rate	Total_Conducted	Total_Passes	Total_Pass rate	year
	All	All	All	All	All	All	All	All	All	All	All	All
18450	Winchester	24	58	27	46.55172	52	18	34.61538	110	45	40.90909	2008
18451	Winchester	25	26	12	46.15385	43	18	41.86047	69	30	43.47826	2008
18452	Wood Green	17	225	122	54.22222	146	69	47.26027	371	191	51.48248	2008
18453	Wood Green	18	326	145	44.47853	301	124	41.19601	627	269	42.90271	2008
18454	Wood Green	19	279	122	43.72760	321	137	42.67913	600	259	43.16667	2008
18455	Wood Green	20	216	90	41.66667	319	120	37.61755	535	210	39.25234	2008
18456	Wood Green	21	218	109	50.00000	266	107	40.22556	484	216	44.62810	2008

Table 1 Data table after preprocessing

The final dataset that was used for analysis consisted of a table of 22060 rows that consisted of observations, organized into 12 columns, which contained test centre location, age, car pass rates for male and female genders, and year. This table was then used to subset two new tables, filtered by location, so as to compare between XYZ's given choice of cities, Rochdale and Wood Green.

Distribution of Pass Rates

The data of both the locations were analyzed visually to explore the dataset.



Graph 1 and 2: Density plots of Male pass rates

When the Total and Male Pass Rates for both locations were plotted, we find that the pass rates follow an approximate bell curve, showing normal distribution around respective means, for their Pass rate densities. Thus, we assume that the pass rate data follows a normal distribution.

Methodology

Mean

XYZ's expected passing rate at Rochdale, which is nearest to his home and at Wood Green, which is nearest to the LSE was found using the mean value of the Male Passing Rate at these locations, for all years.

The mean of the male passing rate is the sum of the male passing values (x) for each year divided by the total number of items in the sample (n).

$$\bar{X} = \frac{\sum x}{n}$$

From the plots of the male pass rates, as shown in Graphs 1 and 2, we see that they follow a normal distribution. We have also assumed normal distribution. Thus, the mean value will give us an accurate value for the expected pass rate for both Rochdale and Wood Green.

Probability of Passing using Naïve Bayes

The method used to statistically determine which test centre location would give XYZ, the best chance to pass his test, was found by comparing the probability of his pass rate at both locations using Naïve Bayes theorem. The Naive Bayes formula used is:

P(outcome | evidence) = P(outcome) * P(evidence | outcome) / P(evidence)

Where the outcome is the Passing of XYZ at a given location and evidence is the age and gender of XYZ. Using this formula, we can calculate the probability of XYZ's pass rate for Rochdale and Wood Green, which can be then compared to determine the location which gives the higher probability of XYZ passing, given that he is a male and that his age is 22.

We have used the Bayes theorem to predict the probability of passing because it is a simple and natural probabilistic method that can be used to compute the classification probabilities, assuming all attributes are independent from each other. In our dataset, this is a valid assumption, as Age and Gender are the attributes used, and they are naturally independent of each other. Year is not used as an attribute in Naïve Bayes method because it is not a categorical variable. The correlation between year and pass rates is also found to be insignificant. Naïve Bayes performs well in case of categorical input variables, which is true in our case.

ASSUMPTIONS

For the given dataset, a few assumptions were made in order use the two methods mentioned above. Firstly, as mentioned before, we have assumed that the total and male pass rates are normally distributed, which we inferred from the graphs 1 to 2.

We then assumed that the categorical variables Age and Gender of the test takers at both location were independent of each other. This is a valid assumption as they are naturally independent of each other over a general population. This assumption allows us to calculate the probability of the evidence by multiplying the individual probabilities of each piece of evidence occurring together using the simple multiplication rule for independent events.

We have also assumed that each location is independent of each other, and the data provided for each location that are complete, are accurate.

RELEVANT R- CODE

LOADING LIBRARIES

#Firstly, two libraries were loaded, readxl and ggpubr, along with the ones applied by default. Readxl #was used in importing the data file.

```
library(readxl)
library("ggpubr")
```

LOADING XYZ PROFILE

```
# Replace the number below by your LSE ID
ID = 20176****
# Then copy XYZprofile.r into your R working directory
source ("XYZprofile.r")
# Now run the function XYZprofile with argument ID
XYZprofile(ID)

## The profile of XYZ:
## - Age:  22
## - Gender:  Male
## - Home address:  Rochdale (Manchester)
```

IMPORTING, PROCESSING AND CLEANING OF DATA

The data was cleaned and processed to give us two tables, one for Rochdale and the other for Wood Green.

```
#cleaning the test data data
year=2014:2007
data=read_excel("dvsa1203.xlsx",sheet=1,col_names = T,col_types ="numeric" ,skip = 6)

data2 <- read_excel("dvsa1203.xlsx",sheet=1,col_names = T,skip = 6)
data$year <- rep(year[1],nrow(data))
#importing all the excel sheets
for (i in 2:8){
 y1=read_excel("dvsa1203.xlsx",sheet=i,col_names = T,col_types ="numeric" ,skip = 6)
 y2=read_excel("dvsa1203.xlsx",sheet=i,col_names = T ,skip = 6)
 y1$year <- rep(year[i],nrow(y1))
 data=rbind(data,y1)
data2=rbind(data2,y2)
}

data$Location <- data2$Location
#clearing empty rows
data<-data[complete.cases(data[ , 2:9]),]
```

```
#creating subsets for both cities
rochdale=data[data$Location =="Rochdale",2:12]
wood_green=data[data$Location =="Wood Green",2:12]
```

	Age	Male_Conducted	Male_Passes	Male_Pass_rate	Female_Conducted	Female_Passes	Female_Pass_rate	Total_Conducted	Total_Passes	Total_Pass rate	year
1	17	381	191	50.13123	339	165	48.67257	720	356	49.44444	2014
2	18	358	176	49.16201	388	165	42.52577	746	341	45.71046	2014
3	19	250	111	44.40000	273	112	41.02564	523	223	42.63862	2014
4	20	178	87	48.87640	217	73	33.64055	395	160	40.50633	2014
5	21	182	82	45.05495	162	70	43.20988	344	152	44.18605	2014
6	22	146	69	47.26027	137	63	45.98540	283	132	46.64311	2014
7	23	95	46	48.42105	130	49	37.69231	225	95	42.22222	2014
8	24	93	50	53.76344	115	49	42.60870	208	99	47.59615	2014
9	25	53	31	58.49057	85	27	31.76471	138	58	42.02899	2014
10	17	329	168	51.06383	336	161	47.91667	665	329	49.47368	2013
11	18	419	190	45.34606	432	177	40.97222	851	367	43.12573	2013
12	19	246	115	46.74797	274	114	41.60584	520	229	44.03846	2013

Table 2: Snippet of subset Rochdale

FINDING EXPECTED PASS RATES AT BOTH CITIES

Expected pass rate of XYZ at Rochdale

```
#finding mean of pass rate for xyz at hometown
mean(rochdale$Male_Pass_rate[rochdale$Age == 22])
```
[1] 45.36534

Expected Pass rate of XYZ at Wood Green

```
#finding mean of pass rate for xyz at LSE
mean(wood_green$Male_Pass_rate[wood_green$Age == 22])
```
[1] 42.31974

CALCULATING PROBABILITY OF PASS RATE OF XYZ AT BOTH LOCATIONS USING NAÏVE BAYES

To calculate the probability of passing, likelihood tables of passing were made for the Age and Gender variables.

```
#creating likelihood tables by age
#for rochdale
age.rc=aggregate(.~Age,rochdale , FUN=sum)
age.rc$`Total_Pass rate` = age.rc$Total_Passes/age.rc$Total_Conducted
age.rc$Female_Pass_rate = age.rc$Female_Passes/age.rc$Female_Conducted
age.rc$Male_Pass_rate = age.rc$Male_Passes/age.rc$Male_Conducted
```

	Age	Male_Conducted	Male_Passes	Male_Pass_rate	Female_Conducted	Female_Passes	Female_Pass_rate	Total_Conducted	Total_Passes	Total_Pass rate	year
1	17	4833	2345	0.4852059	3754	1671	0.4451252	8587	4016	0.4676837	16084
2	18	4010	1767	0.4406484	4602	1736	0.3772273	8612	3503	0.4067580	16084
3	19	2187	949	0.4339278	2766	1012	0.3658713	4953	1961	0.3959217	16084
4	20	1478	678	0.4587280	1820	692	0.3802198	3298	1370	0.4154033	16084
5	21	1272	584	0.4591195	1720	606	0.3523256	2992	1190	0.3977273	16084
6	22	1167	528	0.4524422	1346	522	0.3878158	2513	1050	0.4178273	16084
7	23	1053	462	0.4387464	1212	442	0.3646865	2265	904	0.3991170	16084
8	24	795	363	0.4566038	1136	403	0.3547535	1931	766	0.3966857	16084
9	25	898	416	0.4632517	1035	319	0.3082126	1933	735	0.3802380	16084

Table 3: Likelihood table by age for Rochdale

```
#for wood green
age.wg=aggregate(.~Age,wood_green , FUN=sum)
age.wg$`Total_Pass rate` = age.wg$Total_Passes/age.wg$Total_Conducted
age.wg$Female_Pass_rate = age.wg$Female_Passes/age.wg$Female_Conducted
age.wg$Male_Pass_rate = age.wg$Male_Passes/age.wg$Male_Conducted
```

	Age	Male_Conducted	Male_Passes	Male_Pass_rate	Female_Conducted	Female_Passes	Female_Pass_rate	Total_Conducted	Total_Passes	Total_Pass rate	year
1	17	1518	707	0.4657444	892	386	0.4327354	2410	1093	0.4535270	16084
2	18	2277	983	0.4317084	2160	817	0.3782407	4437	1800	0.4056795	16084
3	19	2059	900	0.4371054	2376	901	0.3792088	4435	1801	0.4060879	16084
4	20	1675	708	0.4226866	2067	790	0.3821964	3742	1498	0.4003207	16084
5	21	1542	677	0.4390402	1860	694	0.3731183	3402	1371	0.4029982	16084
6	22	1496	631	0.4217914	1730	649	0.3751445	3226	1280	0.3967762	16084
7	23	1352	582	0.4304734	1769	617	0.3487846	3121	1199	0.3841717	16084
8	24	1125	547	0.4862222	1758	657	0.3737201	2883	1204	0.4176205	16084
9	25	1283	561	0.4372564	1792	581	0.3242188	3075	1142	0.3713821	16084

Table 4: Likelihood table by age for Wood Green

```
#taking sum of columns for both cities
sum.rc=colSums(rochdale[2:9])
sum.wg=colSums(wood_green[2:9])
#likelihood table for gender
#for rochdale
sum.rc["Total_Pass rate"] = sum.rc["Total_Passes"]/sum.rc["Total_Conducted"]
sum.rc["Female_Pass_rate"] = sum.rc["Female_Passes"]/sum.rc["Female_Conducted"]
sum.rc["Male_Pass_rate"] = sum.rc["Male_Passes"]/sum.rc["Male_Conducted"]
```

	Male_Conducted	Male_Passes	Male_Pass_rate	Female_Conducted	Female_Passes	Female_Pass_rate	Total_Conducted	Total_Passes	Total_Pass rate
1	17693	8092	0.457356	19391	7403	0.3817751	37084	15495	0.4178352

```
#for wood green
sum.wg["Total_Pass rate"] = sum.wg["Total_Passes"]/sum.wg["Total_Conducted"]
sum.wg["Female_Pass_rate"] = sum.wg["Female_Passes"]/sum.wg["Female_Conducted"]
sum.wg["Male_Pass_rate"] = sum.wg["Male_Passes"]/sum.wg["Male_Conducted"]
```

	Male_Conducted	Male_Passes	Male_Pass_rate	Female_Conducted	Female_Passes	Female_Pass_rate	Total_Conducted	Total_Passes	Total_Pass rate
1	14327	6296	0.43945	16404	6092	0.3713728	30731	12388	0.4031109

```
#probabilities required for bayes formula
#probability of being age 22 for both cities
pr.age.rc=age.rc$Total_Conducted[age.rc$Age==22]/sum.rc[["Total_Conducted"]]
pr.age.wg=age.wg$Total_Conducted[age.wg$Age==22]/sum.wg[["Total_Conducted"]]
#probability of being male
pr.male.rc=sum.rc[["Male_Conducted"]]/sum.rc[["Total_Conducted"]]
pr.male.wg=sum.wg[["Male_Conducted"]]/sum.wg[["Total_Conducted"]]
#probability of being age 22 given passed for both cities
pr.pass.age.rc=age.rc$Total_Passes[age.rc$Age == 22]/sum.rc[["Total_Passes"]]
pr.pass.age.wg=age.wg$Total_Passes[age.wg$Age == 22]/sum.wg[["Total_Passes"]]
```

```
#probability of being male given passed
pr.pass.male.rc=sum.rc[["Male_Passes"]]/sum.rc[["Total_Passes"]]
pr.pass.male.wg=sum.wg[["Male_Passes"]]/sum.wg[["Total_Passes"]]

#finding the bayesian probability of XYZ passing in hometown, given Age and Gender at both locations
#for rochdale
prob.pass.rc=(pr.pass.age.rc * pr.pass.male.rc * sum.rc[["Total_Pass rate"]])/(pr.age.rc*pr.male.rc)
#for wood green
prob.pass.wg=(pr.pass.age.wg * pr.pass.male.wg * sum.wg[["Total_Pass
rate"]])/(pr.age.wg*pr.male.wg)
```

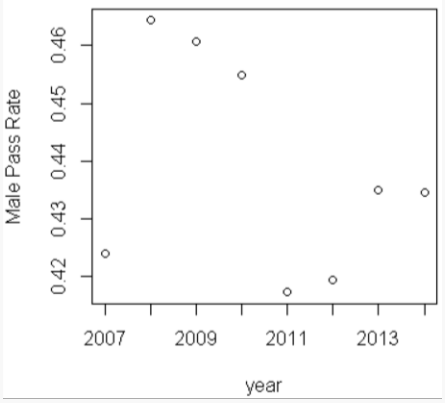
prob.pass.rc = 0.4573474

prob.pass.wg = 0.4325443

prob.pass.rc/(prob.pass.rc+prob.pass.wg)

[1] 0.513936

#correlation between year and male pass rate for wood green

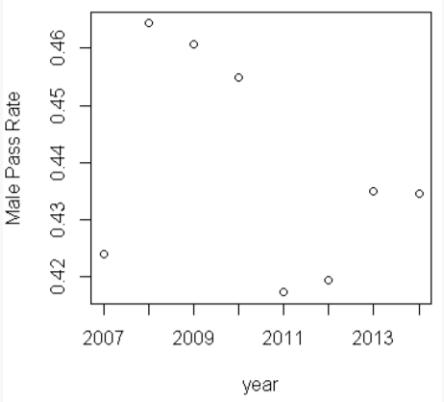


Graph 3: Male pass rate vs Year for Wood Green

cor(b.wg$year,b.wg$Male_Pass_rate)

[1] -0.3627893

#correlation between year and male pass rate for Rochdale



Graph 4: Male pass rate vs Year for Rochdale

cor(b.rc$year,b.rc$Male_Pass_rate)

[1] 0.6520871

RESULTS

We observe that the XYZ's expected passing rate which is the mean of the pass rate of a male of age 22 at Rochdale is 45.4 % and that for Wood Green is 42.3%.

We also see that, using Naïve Bayes, the probability of XYZ passing the test in his hometown Rochdale is 45.7% and the probability of passing the test near LSE at Wood Green is 43.2 %. Rochdale is thus relatively slightly more favorable for XYZ. The correlation between year and the male pass rate were plotted and calculated using the cor() function and observed that the year and male pass rates for both locations were not correlated. Thus, year was not a significant predictor of pass rate.

CONCLUSION

Finally, we calculated that XYZ has an expected passing rate of 45.4% at Rochdale and an expected passing rate of 42.3% at Wood Green. This already indicates that Rochdale gives XYZ a slightly better chance than Wood Green. This was confirmed, when we found that the probability of XYZ passing at Rochdale was better by using Naïve Bayes, which found the probability of passing as 45.7% at Rochdale and 43.2 % at Wood Green. Thus, we find that Rochdale is better for XYZ with an edge of about 3%. Although, XYZ's chances are not significantly improved by changing locations from Rochdale to Wood Green, it would be slightly better to take the test at Rochdale, as per convenience. This means that, if XYZ is already near LSE, it would not be worthwhile to travel back to Rochdale as his chances will only be slightly improved.

The benefits of comparing the two locations on the basis of XYZ's probability of passing using Naïve Bayes is that, Naïve Bayes is easy and fast to predict the probability of an outcome, even though we have categorical character predictors, such as gender. A limitation of this method is that naivety results in probabilities that are not entirely mathematically correct but they are a good approximation and adequate for the purposes of classification, which in our case, is the choice between the locations. Another limitation of Naïve Bayes is the assumption of independent predictors. In our case we assumed that gender and age are independent, although it could be possible that they are not completely independent of each other, in this dataset.

REFERENCES

James, G., Witten, D., Hastie, T., Tibshirani, R. *An Introduction to Statistical Learning*. Springer Publishing Company, 2013.