National College of Ireland

Project Submission Sheet – 2019/2020

Student Name:	Antonio Caruso		
Student ID:	19203608		
Programme:	Higher Diploma in Science in Data Analytics	**Year:**	2020
Module:	Advanced Business Data Analysis		
Lecturer:	Theo Mendonca		
Submission Due Date:	26/07/2020		
Project Title:			
Word Count: 2003			

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: ………

Date:

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Contents

1 Introduction

This paper is a report on parametric and non parametric statistical tests applied to a dataset collected from the Central Statistical office of Ireland (Central Statistics Office, 2016)

2 First Report- Comparison of methods of transports to work in County Dublin

The first reports analyses a comparison of transports to work in county Dublin. For this test, one independent variable (County Dublin) and two dependent variables corresponding to the number of people who either commuted by public transport or by car (Central Statistics Office, 2016)

2.1 Research Question: Is there any statistical difference between the two groups of commuters?

In order to respond to the question, the T-test was initially considered to compare the means of the two independent groups (commuters and car drivers to work).

2.2 Specification of the null and alternate hypotheses and Alpha level

H_0: $\mu_{(transport)} = \mu_{(car)}$ **There is no difference between the means of the two groups**

H1: $\mu_{(Transport)} \neq \mu_{(car)}$ **There is difference between the means of the two groups**

$\alpha = 0.05$

2.3 Test for normality

in order to understand whether to use a parametric or non-parametric test, the two distributions were tested for normality. The initial assumption was that the dependent variables are normally distributed and that a parametric variable should be applied.

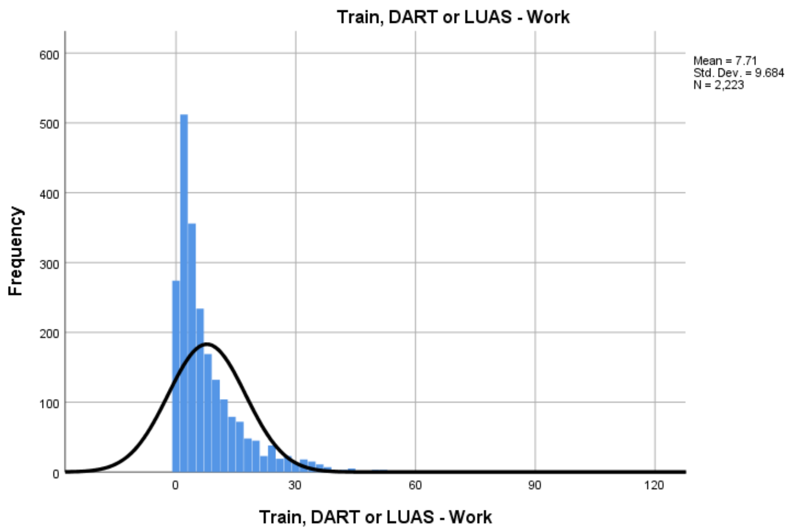
Statistics

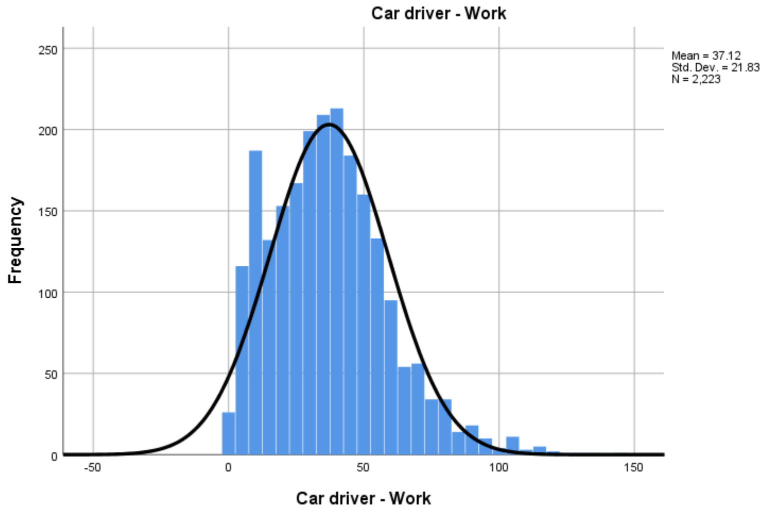
		Train, DART or LUAS - Work	Car driver - Work
N	Valid	2223	2223
	Missing	1	1
Mean		7.71	37.12
Std. Error of Mean		.205	.463
Median		4.00	36.00
Mode		0	44
Std. Deviation		9.684	21.830
Variance		93.773	476.555
Skewness		2.861	.763
Std. Error of Skewness		.052	.052
Kurtosis		13.749	.942
Std. Error of Kurtosis		.104	.104
Range		107	143
Sum		17139	82516
Percentiles	25	2.00	21.00
	50	4.00	36.00
	75	10.00	50.00

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Transport Method	Statistic	df	Sig.	Statistic	df	Sig.
Commuters	Car driv	.048	2223	.000	.962	2223	.000
	Train, D	.213	2223	.000	.723	2223	.000

a. Lilliefors Significance Correction





The above histograms obtained from the frequency distribution of the two variables of commuters to works and Car drivers to work shows a positively skewed distribution with a skewness of 2.86 for commuters and a normal distribution for Car Driver-Work. The Shapiro-Wilk test for normality (McClave & Sincich, 2017) resulted in 0,0001 < p-value 0.05, therefore the null hypothesis that data are normally distributed was rejected. A non-parametric test was then necessary to be applied, thus Man-Whitney U test (McClave & Sincich, 2017) was chosen to determine whether there is a significant difference between the two variables since the commuters are not normally distributed.

2.4 New Hypothesis

- Null Hypothesis (H_0)

The distribution of commuters is the same across those who take transports and those who drive

- Alternate Hypothesis (H_1)

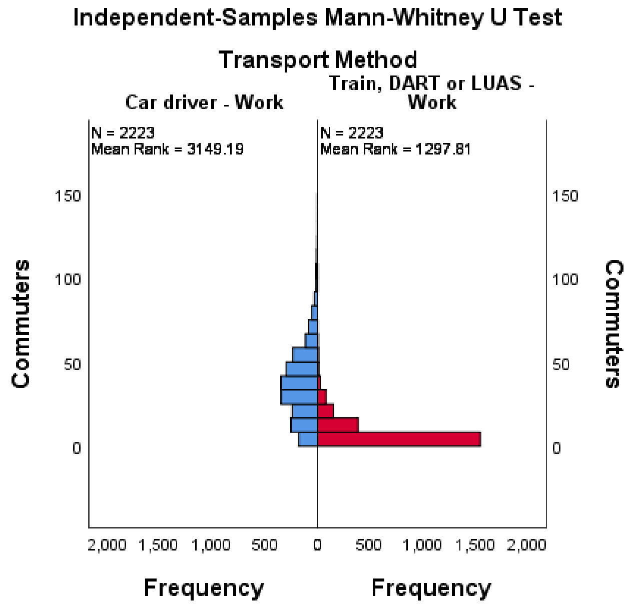
  The distribution of commuters is not the same across those who take transports and those who drive

  α: 0.05

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Commuters is the same across categories of Transport Method.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.



2.5 Interpretation

The first column on the summary table expressed the null hypothesis in terms of distribution of commuters for both transports and car. The Sig. displays the statistical significance of the Mann-Whitney U test as the p-value (McClave & Sincich, 2017). Because the obtained p-value 0.000001 < 0.5 significance level chose in the beginning, The Null Hypothesis was rejected with conclusion that the distribution of commuters is the same across transport commuters and car drivers.

3 Second Report – Comparison of presence of three Ethnical races in county Carlow and Cavan

Would a particular county affect the decision of living there of three different ethnical groups?"

The second reports refers to a comparison of three ethnical races in county Carlow and Cavan. Initial plan was to apply a two- way ANOVA (reference) to compare the presence of three difference ethnical races as dependent variable in two counties Carlow and Cavan as independent variable (Central Statistics Office, 2016).

3.1 Specification of the initial 3 hypotheses and alpha level

1. **H_1:** All the ethnic groups have equal presence across the counties
2. **H_2:** Both the county groups affect presence of ethnic groups
3. **H3:** The ethnicity and County factors are independent or interaction effect is not presence

α***: 0.05***

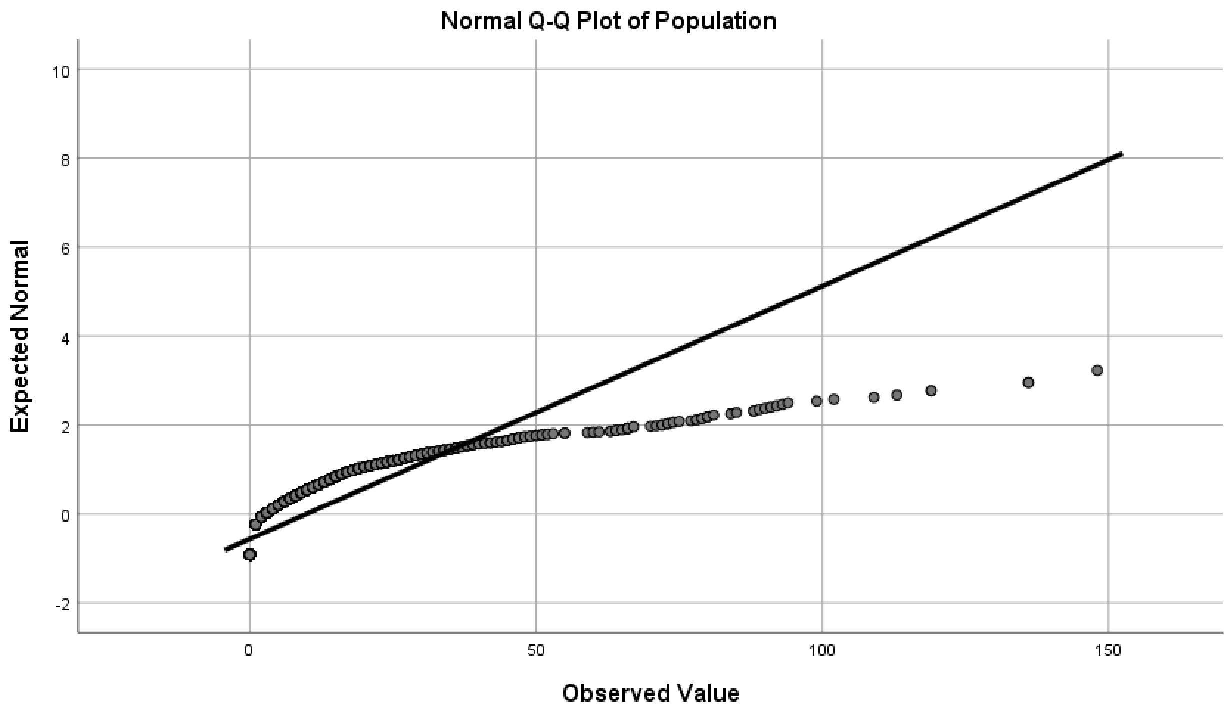
3.2 Test for normality

The Shapiro-Wilk test was used to test the data for normality (McClave & Sincich, 2017)

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	County	Statistic	df	Sig.	Statistic	df	Sig.
Population	Carlow	.288	639	.000	.589	639	.000
	Cavan	.286	969	.000	.608	969	.000

a. Lilliefors Significance Correction



The result shoed clearly that the points of the two sets of quantiles for the distribution of Carlow and Cavan created against one another with the QQ plot did not form a straight line that should be expected. As the level of Significance obtained from the Shapiro-Wilk was 0.001 < 0.05, the null hypothesis that the values are normally distributed was rejected.

Since the three groups are were normally distributed, it was necessary to apply a non parametric test**.** Not knowing a non-parametric version of the two-way anova, The Kruskas Wallis H test (McClave & Sincich, 2017) was applied to test new hypotheses as below;

3.3 New hypothesis and question

Is there a significant difference between the groups of white non Irish, black/black Irish and Asian/Asian Irish located between County Carlow and Navan?

- *H_0: population medians are equal.*
- *H_1: population medians are not equal.*
- α: 0.05

3.4 Interpretation of results and Post Hoc test

```
        Kruskal-Wallis rank sum test

data:  Groups
Kruskal-Wallis chi-squared = 800.63, df = 2, p-value
< 2.2e-16
```

The very High chi-squared shows that there is not relationships among the three groups and that the data do not fit very well. A p value obtained of 0.00000000000000022 < 0.05 meant that significant difference between the groups exists. The post hoc testing was conducted to understand which groups are different from the others.

Pairwise Comparisons of Ethnicity

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Black or Black Irish-Asian or Asian Irish	62.638	27.674	2.263	.024	.071
Black or Black Irish-Other White	-707.282	27.674	-25.558	.000	.000
Asian or Asian Irish-Other White	-644.644	27.674	-23.294	.000	.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

After running the post hoc test with a level of significance adjusted by the Bonferroni correction, we can see that the presence the Black or black Irish and other white non Irish is significantly different from each other, as well as the Asian or Asian Irish and white non Irish are significantly different from each other, whereas black or black Irish and Asian or Asian Irish are not. Then H0 Hypothesis that the population medians are equal was rejected in favour of the alternative hypothesis.

4 Third Report- Comparison of Leaving time to work, school or college

The third report analyses a comparison of leaving time to work, school or college between southern and eastern midland areas (Central Statistics Office, 2016) with the goal of answering the following research question;

Research question: Do Southern people leaving time to work, school or college in southern area differ significantly from people from eastern midland areas?

In order to answer to the research question, the non-parametric test of Chi-Squared of independence (McClave & Sincich, 2017)was considered to conduct a statistical test to answer the research question.

4.1 Specification of the null and alternate hypotheses and Alpha level

Hypothesis

H0: Leaving time and areas are independent

H1: Leaving time and areas are not independent

α: 0.05

4.2 Test for normality

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Southern	.336	3	.	.856	3	.256
eastern midlands	.362	3	.	.804	3	.123

a. Lilliefors Significance Correction

The Chi squared test of independence would apply only in case of non normal distribution. To verify that, Shapiro Test was applied to check whether the distribution was normal. Both the distribution of data for Southern and eastern midlands resulted not normal, and the null hypothesis that $p > 0.05$ is rejected as Southern presents a sig. level of 0.2 and eastern midlands 0.12. As the data are not normally distributed, the Chi test squared for independence can be applied as planned.

4.3 Test Execution

LEAVING TIME BY AREA

	before 6.30	7:31 - 8:00	after 9:30	*Row Total*
southern	99	507	158	764
Eastern Midland	93	312	76	481
Column Total	*192*	*819*	*234*	1245

$$E_{r,c} = \frac{(n_r * n_c)}{n}$$

	before 6.30	7:31 - 8:00	after 9:30	
southern(exp)	117.8216867	502.583133	143.5951807	
(O-E)	-18.82168675	4.41686747	14.40481928	
(O-E)2	354.255892	19.5087182	207.4988184	
(O-E)2/E	3.006712107	0.0388169	1.44502634	**4.490555345**

	before 6.30	7:31 - 8:00	after 9:30	row total
Eastern Midland(exp)	74.17831325	316.416867	90.40481928	
(O-E)	18.82168675	-4.4168675	-14.40481928	
(O-E)2	354.255892	19.5087182	207.4988184	
(O-E)2/E	4.775733991	0.06165511	2.295218552	**7.132607658**

$$\chi^2 = \sum \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}$$

Chi	**11.623163**
DF	2
Crit	5.991
Chi stat > Chi crit	**Reject H0**

$$DF = (r - 1) * (c - 1)$$

DF	P 0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

The test was conducted at an alpha value of 0.05. The first step involved the calculation of two expected frequency counts for the two different areas in each of the three groups of leaving time. The formula used referred to the total number of the rows multiplied by the total number of columns divided by the grand total. As the Chi Square formula equals to the sum of total observed values minus the expected values squared, divided by the expected values, the O-E, O-E and O-E2/E were calculated. The total for the two areas resulted in 4.49 for Southern area and 7.13 for Eastern Midland. The final Chi value corresponding to the sum of the two numbers, resulted in CHI = 11.62.

In order to determine if this was a significant chi statistics, the Chi distribution table was observed. The degree of freedom DF corresponded to 2 based on the formula. For two degrees of freedom at alpha level of 0.5, the Critical value corresponds to 5.991

4.4 Interpretation of the output

As the Chi stat 11.63 > Chi Crit 5.991, the null hypothesis H0 that the leaving time and the areas are independent was rejected favour of the alternative hypothesis, which states that there is a relationship between the time when people leave their home and the area from which they leave in the morning to go either to work, school or college. This answers the research question in regards to difference in the two areas about the time leaving home.

5 Fourth Report – Comparison of three different Journey times to work in the small area of Ardee (County Louth)

The fourth report represents a comparison of journey times to work in the small area of Ardee located in County Louth (Central Statistics Office, 2016).

The research question is: is there a significant difference in journey times to work experienced by people that live in the small area of Ardee?

In order to answer the research question, the parametric test One-way Anova was considered the best to compare the means of two or more groups (McClave & Sincich, 2017). For this test, three different journey times were analysed;

1- Under 15 minutes
2- ¼ hour – under ½ hour
3- 1 hour – under ½ hour

The interest in this test lies in whether the difference between the travelled journey of people in Ardee is significantly different than expected in random variation within groups.

5.1 Specification of the null and alternate hypothesis and Alpha level

H_0: $\mu_{und15} = \mu_{1/4\text{-}und1/2} = \mu_{1h\text{-}und1/2}$ There is no difference between the means of the three groups

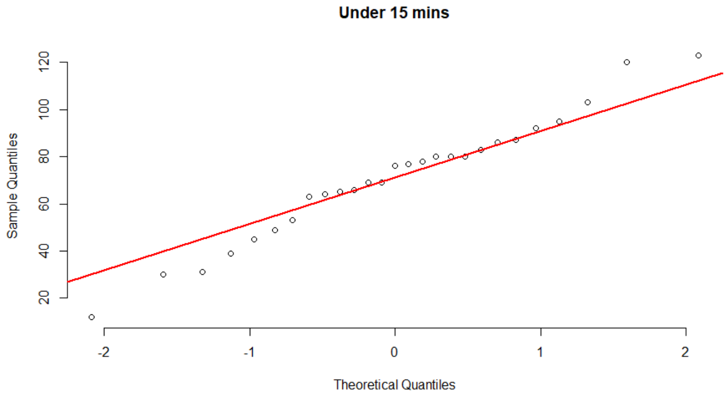
$H1 = \mu_{und15} \neq \mu_{1/4\text{-}und1/2} \neq \mu_{1h\text{-}und1/2}$ There is difference between the means of the three groups

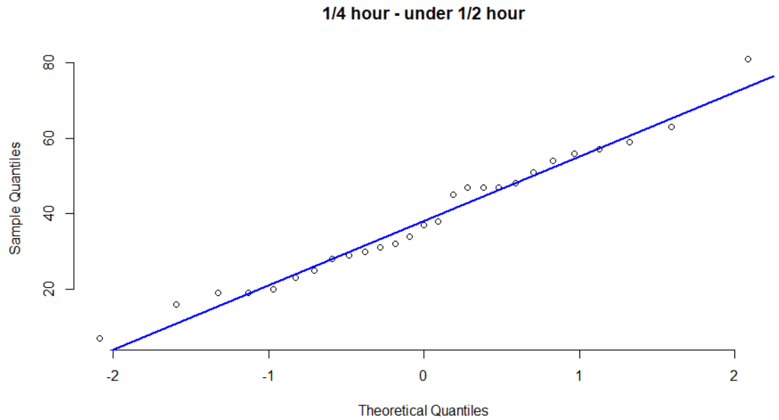
α: 0.05

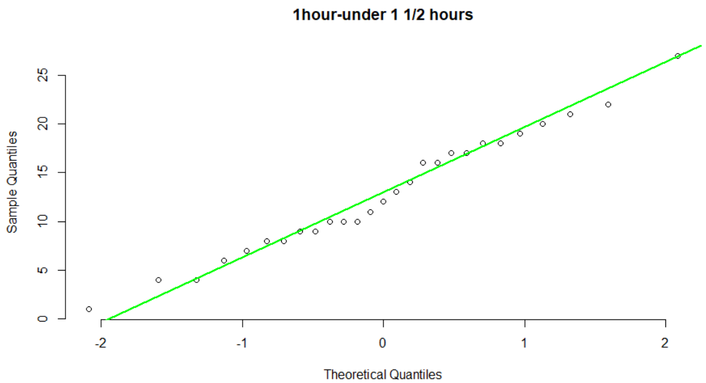
5.2 Test for Normality

```
> shapiro.test(Ardee$`under 15 minutes`)

        Shapiro-Wilk normality test

data:  Ardee$`under 15 minutes`
W = 0.9775, p-value = 0.8021

> shapiro.test(Ardee$`1/4 hour - under 1/2 hour`)

        Shapiro-Wilk normality test

data:  Ardee$`1/4 hour - under 1/2 hour`
W = 0.97649, p-value = 0.776

> shapiro.test(Ardee$`1 hour - under 1 1/2 hours`)

        Shapiro-Wilk normality test

data:  Ardee$`1 hour - under 1 1/2 hours`
W = 0.9816, p-value = 0.8966
```

The Shapiro test (McClave & Sincich, 2017)was applied to evaluate the normal distribution of the three groups. The H0 shows that the data are normally distributed across the three independent groups. The "under 15 minutes group" presents a p-value of 0.80 > 0.05, the second group displays a p value of 0.776 >0.05, and the third group has 0.89 >0.05. The null hypothesis that the data are normally distributed can be accepted and the one-way Anova test could be applied.



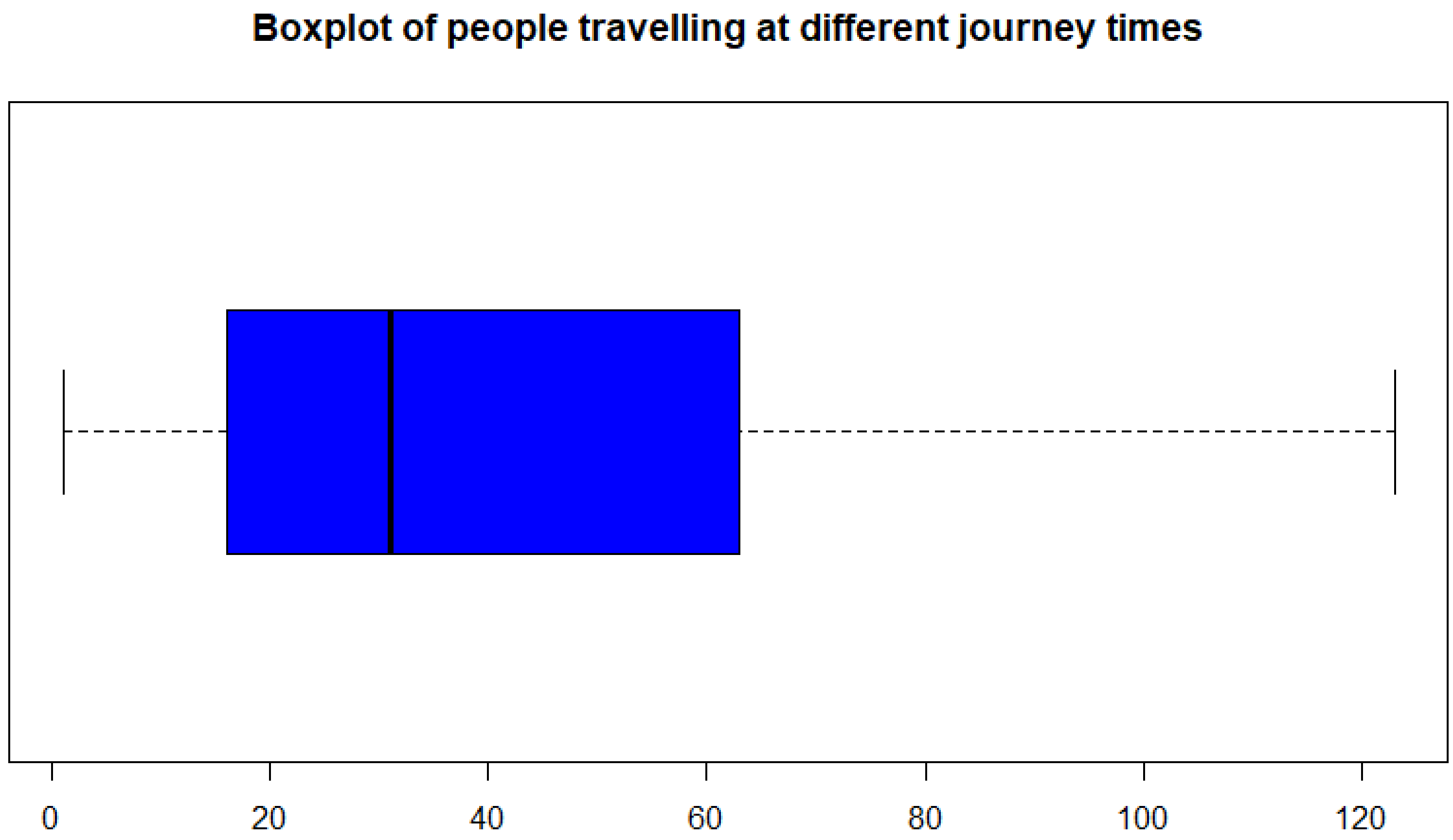




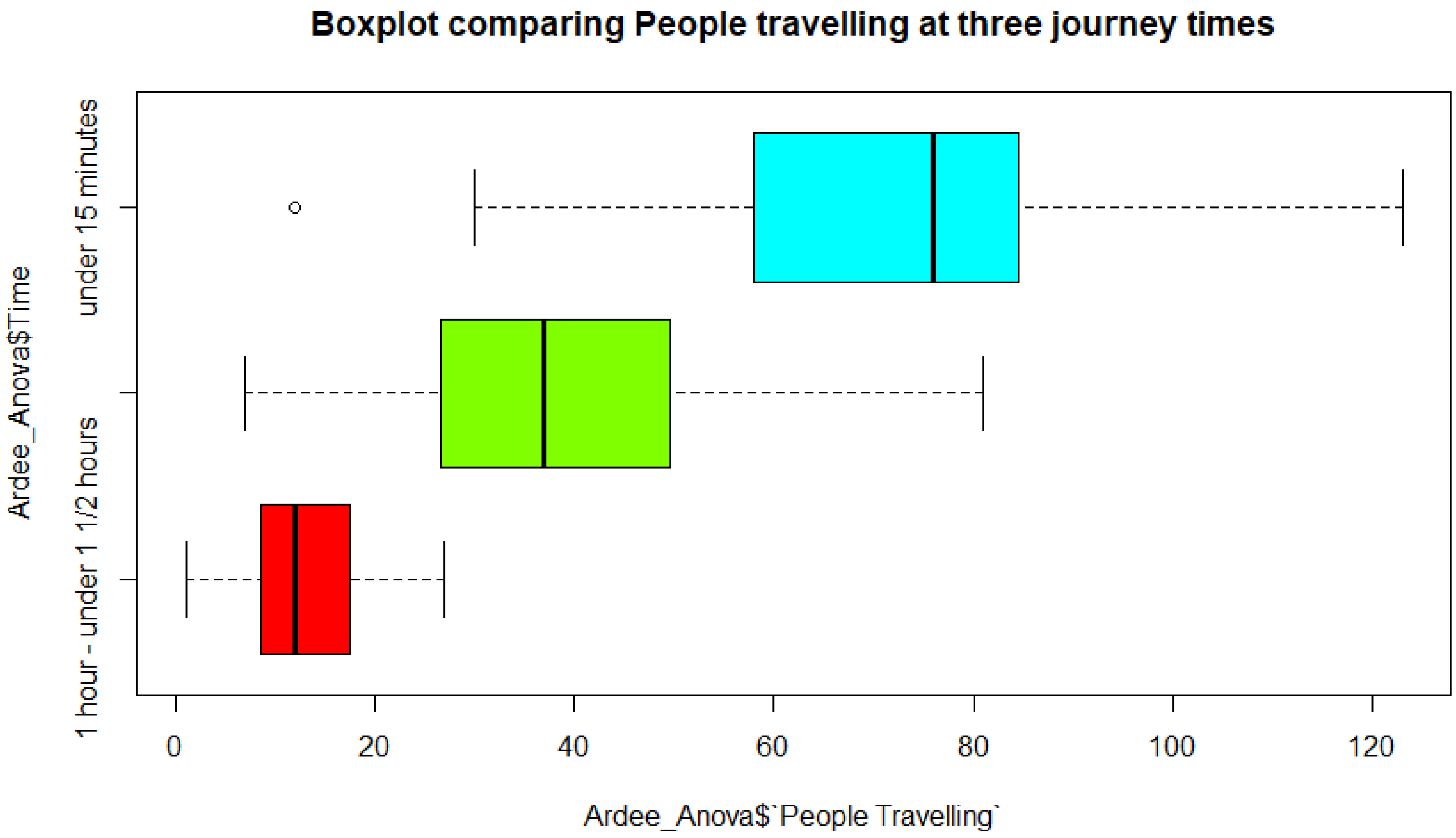
As all the points form a line that is straight, it can be visually confirmed what expressed by the shapiro test that the three distributions are normal.

5.3 One way Anova test execution

The boxplot below helped understand the data before executing the test;



The box plot doesn't show a perfect symmetry and displays the max number of people that travelled at the identified times is around 122, and the median around 30 people that travelled across those journey time



Each group presented 27 observations. The more clustered the value are, the smaller the standard deviation would be (McClave & Sincich, 2017). According to the plot, Under 15 minutes seems to be the most common journey time for people in Ardee, with a median of 75 people travelling within that short timeframe. As the outlier is far from the range, it might be an error and should not happen very often. The lowest median refer to the people that travel for 1hr/under 1h ½ hours with 15 people, as potentially expected in a small area.

```
> tapply(Ardee_Anova$`People Travelling`, Ardee_Anova$Time, var)
1 hour - under 1 1/2 hours  1/4 hour - under 1/2 hour          under 15 minutes
                  39.66952                  293.78063                 675.99430
```

With the lowest variance square of the standard deviation, Under 15 minutes has the highest uncertainty and 1 hour or less than 1:30 has the lowest.

```
> Ardee_Anova.aov <- aov(`People Travelling`~ Time, Ardee_Anova)
> summary(Ardee_Anova.aov)
            Df Sum Sq Mean Sq F value Pr(>F)    
Time         2  45721   22861   67.94 <2e-16 ***
Residuals   78  26246     336                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

The ANOVA formula shows an F (2,79) = 67.93, with 2 and 79 being within groups. As the p value obtained is 0.00000000000000022 < 0.05, the Null hypothesis was rejected in favour of alternative hypothesis $H1 = \mu_{und15} \neq \mu_{1/4-und1/2} \neq \mu_{1h-und1/2}$ that there is a statistical different in at least one journey time. A Post Hoc test was conducted with the following to adjust the p values. Having the same number of observations per group and being less conservative than Bonferroni (Chen, et al., 2017), the Tukey Post hoc test was chosen and delivered the results;

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = `People Travelling` ~ Time, data = Ardee_Anova)

$Time
                                                       diff      lwr      upr   p adj
1/4 hour - under 1/2 hour-1 hour - under 1 1/2 hours 25.77778 13.84951 37.70604 5.4e-06
under 15 minutes-1 hour - under 1 1/2 hours          58.07407 46.14581 70.00234 0.0e+00
under 15 minutes-1/4 hour - under 1/2 hour           32.29630 20.36803 44.22456 0.0e+00
```

The conclusion is that all the times are statistically different as the p adjusted values is < 0.5 for all the possible combinations

6 Fifth Report – Difference between two samples of people owning a car living in two neighbour urban areas

Research question: Is there any difference between the two urban areas that determine the number of people owning 1 car?

The final report is about a difference of two related samples of people owning a car in the urban area 1 and urban area 1 of the small area of Dungarvan in county Waterford (Central Statistics Office, 2016). The scope was to understand whether there was a statistical difference between the two samples. A potential Two-sample T test (McClave & Sincich, 2017) could be applied

6.1 Specification of the null and alternate hypothesis and Alpha level

$H0 : \mu_{urb1} = \mu_{urb2}$

$H1: \mu_{urb1} \neq \mu_{urb2}$

$\alpha : 0.05$

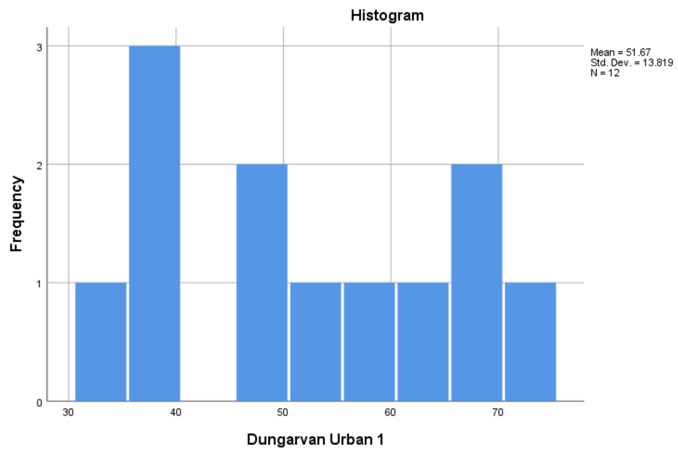
6.2 Test for normality

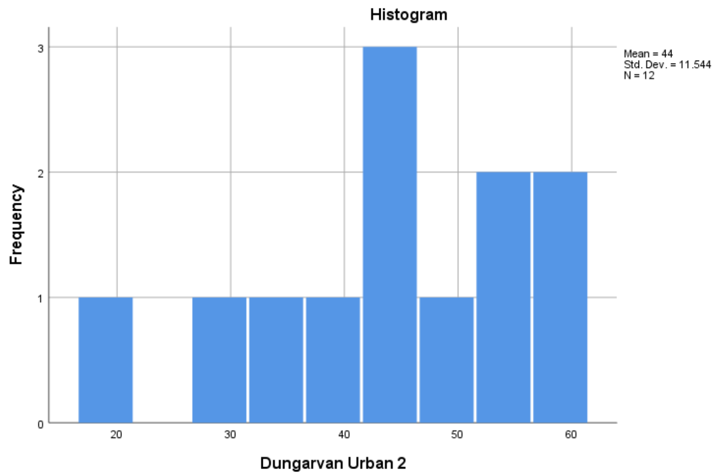
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Dungarvan Urban 1	.189	12	.200*	.914	12	.237
Dungarvan Urban 2	.132	12	.200*	.938	12	.479

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction





With a level of significance for both urban areas of 0.2 and 0.4 < 0.5, the null hypothesis that both data were normally distributed was rejected. Therefore, the non parametric test of Wilcoxon Rank could be applied with the new hypothesis referring to the median rather than mean being a non parametric test as following;

- H0 = the median difference between the pair of observation is zero H_0: $M_{urb1} = M_{urb2}$
- H1 =The median difference between the pair of observations is not zero H_1: $M_{urb1} \neq M_{urb2}$
- α : 0.05

6.3 Interpretation of the test

Dungarvan Urban 1	Dungarvan Urban 2	Differences	| differences|	Sorted |Differences|	Rank	RANK Sum of the Positive Diff	Rank Sum of the Negative Diff
52	30	22	22	3	1	189	21
33	52	-19	19	4	2		
58	36	22	22	6	3		
36	55	-19	19	8	4		
62	47	15	15	15	5		
37	40	-3	3	19	6		
66	58	8	8	19	7		
71	46	25	25	20	8		
37	57	-20	20	22	9		
70	19	51	51	22	10		
49	43	6	6	23	11		
49	45	4	4	25	12		
44		44	44	27	13		
55		55	55	30	14		
30		30	30	37	15		
37		37	37	40	16		
40		40	40	44	17		
23		23	23	44	18		
27		27	27	51	19		
44		44	44	55	20		

The first step consisted in calculating the difference between the sample Urban 1 and Urban 1. Once determined the absolute differences, they were sorted from the smallest to the largest. Subsequently, a rank from the smallest to the largest was created. Finally, the rank for all positive differences (T+) of the highlighted figures, and negative (T-) circled figures were added.

The Rank Sums as follows:

T+ = 189

T- = 21

Wstat = 21 (The smallest was chosen)

n	Two-Tailed Test		One-Tailed Test	
	α = .05	α = .01	α = .05	α = .01
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49
22	65	48	75	55
23	73	54	83	62
24	81	61	91	69
25	89	68	100	76
26	98	75	110	84
27	107	83	119	92
28	116	91	130	101
29	126	100	140	110
30	137	109	151	120

By comparing the value of Wilxocon test statistic to the critical value in the table, considering 20 observations, the obtained value of 21 is statically significant being 21 < 52 of the value in the table;

$W_{stat} = 21$

$\alpha = 0.05$

$n = 20$

$W_{crit} = 52$

$W_{stat} < W_{crit}$

As a conclusion, Rejecting H_0, it can be stated that there is a difference between the two urban areas regarding their condition of owning one cars and it was unlikely to have occurred by chance.

7 References

Central Statistics Office, 2016. *Census 2016 Small Area Population Statistics.* [Online]
Available at: https://www.cso.ie/en/census/census2016reports/census2016smallareapopulationstatistics/
[Accessed 24 July 2020].

Chen, S.-Y., Feng, Z. & Yi, X., 2017. A general introduction to adjustment for multiple comparisons. *Journal of Theoracic Disease,* Volume 9, p. 1725–1729.

McClave, J. & Sincich, T., 2017. *Statistics, Global Edition..* 13th ed. Harlow: Person.