Table of Contents

Advanced Business Data Analytics

An Analysis of Census 2016 Small Area Population Statistics

Ross Currid

x20147147@student.ncirl.ie

***Abstract*— Five Insights have been chosen as part of this analysis of the Census 2016 Small Area Population Statistics (SAP) report. Themes such as unemployment, commute time and home ownership levels are explored. As a corollary, all insights have to some extent been cross referenced with the OSI National Statistics Boundaries 2015 report so as to explore how these themes vary across regions.**

I. Introduction

Statistical Analysis is the cornerstone of data analytics. It is the science upon which data can be interpreted in terms of its statistical significance. In this regard, significance refers to rejection or acceptance of one of two defined hypothesis – H_0 (Null) and H_1 (Alternative) – at a defined level of confidence. With this in mind, all insights have been conducted with a view to not only providing meaningful, interesting insights, but also quantify the ensuing results.

In sections II and III, unemployment by gender and region is explored so as to determine if there is a statistically significant difference in employment levels between genders and across regions. Sections IV and V contain an analysis of commute time alongside a corollary analysis of leaving times for said commutes across regions. Finally, in Section VI, home ownership levels across regions are investigated to determine if, and where, differences in home ownership occur.

In choosing these themes for investigation, current sentiment was considered. Issues such as rent inflation, the perceived gender employment gap and traffic management have been popular discourse as of late.

In conducting analysis of these themes, data was reshaped and reformatted using a variety of R packages, such as Reshape2 (melting data), Dplyr and Tidyr following in the Hadley Wickham vein of gathering 'tidy data' (Wickham, 2014).

II. Unemployment By Region

Both this insight, as well as insight III. were developed as corollaries, for this reason this section (II.) will contain some discussion pertaining to unemployment by gender.

A. Pre-Processing – For both unemployment by region and by gender (III.)

Several columns were chosen within the SAP report, and subsequently subsetted into a new dataframe. Totals for those who are looking after the family home, looking for first time job and unemployed having lost or given up previous job were extracted, as were columns relevant to each gender. 'NUTS3NAME', hereby referred to as 'Region' was extracted so as to relate these columns to their relevant region. Two dataframes were created by subsetting the initial dataframe by columns ending in 'M' or 'F' (denoting Gender). A new column in each dataframe called 'Gender' was created and subsequently filled with either 'Male' or 'Female'. Finally, the two datasets were merged using the rbind function as follows, thus enabling analysis by Gender:

```
Insight2Males <- Insight2[,c(1:3,7)]

Insight2Males <- Insight2Males %>% rowwise %>% rename(LAHF = T8_1_LAHFM,
                                                      ULGUP = T8_1_ULGUPJM,
                                                      LFFJ = T8_1_LFFJM,
                                                      Region = NUTS3NAME)

Insight2Females <- Insight2[,c(4:7)]

Insight2Females <- Insight2Females %>% rowwise %>% rename(LAHF = T8_1_LAHFF,
                                                          ULGUP = T8_1_ULGUPJF,
                                                          LFFJ = T8_1_LFFJF,
                                                          Region = NUTS3NAME)

Insight2Females$Gender <- "Female"
Insight2Males$Gender <- "Male"

Insight2spruced <- rbind(Insight2Females, Insight2Males)
```

B. Summary Statistics for Insights II. And III.

The following summary statistics were gathered:

Gender	Region	LFFJ_Mean	ULGUP_Mean	LAHF_Mean
Female	Border	0.686	5.78	13.2
Female	Dublin	0.883	6.40	14.8
Female	Mid-East	0.832	6.89	18.6
Female	Mid-West	0.715	5.94	15.3
Female	Midlands	0.939	7.74	17.4
Female	South-East	0.774	6.65	16.7
Female	South-West	0.539	4.64	14.7
Female	West	0.588	5.01	12.9
Male	Border	0.839	8.46	1.17
Male	Dublin	1.06	8.16	1.01
Male	Mid-East	1.05	8.91	1.20
Male	Mid-West	0.951	8.42	1.11
Male	Midlands	1.07	10.1	1.27
Male	South-East	1.00	9.62	1.10
Male	South-West	0.700	6.55	0.992
Male	West	0.790	7.42	1.32

Gender	Region	LFFJ_Total	ULGUP_Total	LAHF_Total
Female	Border	1239	10445	23867
Female	Dublin	4309	31257	72357
Female	Mid-East	1952	16153	43578
Female	Mid-West	1375	11423	29393
Female	Midlands	1048	8635	19469
Female	South-East	1302	11185	28065
Female	South-West	1558	13418	42324
Female	West	1174	9993	25756
Male	Border	1516	15286	2117
Male	Dublin	5184	39846	4929
Male	Mid-East	2471	20905	2814
Male	Mid-West	1828	16194	2126
Male	Midlands	1195	11301	1412
Male	South-East	1684	16181	1842
Male	South-West	2023	18932	2867
Male	West	1576	14808	2640

Gender	Region	LFFJ_SD	ULGUP_SD	LAHF_SD
Female	Border	0.955	4.21	6.13
Female	Dublin	1.19	5.77	8.49
Female	Mid-East	1.07	5.11	7.46
Female	Mid-West	0.973	4.55	6.88
Female	Midlands	1.12	5.73	7.10
Female	South-East	1.09	4.77	7.51
Female	South-West	0.830	3.93	7.07
Female	West	0.885	3.97	6.21
Male	Border	1.04	5.79	1.16
Male	Dublin	1.43	7.98	1.15
Male	Mid-East	1.30	6.55	1.19
Male	Mid-West	1.27	6.68	1.15
Male	Midlands	1.33	7.77	1.20
Male	South-East	1.29	6.32	1.14
Male	South-West	0.977	5.32	1.07
Male	West	1.04	5.23	1.31

Gender	Region	LFFJ_Median	ULGUP_Median	LAHF_Median
Female	Border	0	5	12
Female	Dublin	1	5	14
Female	Mid-East	1	6	18
Female	Mid-West	0	5	15
Female	Midlands	1	6	17
Female	South-East	0	6	16
Female	South-West	0	4	14
Female	West	0	4	12
Male	Border	1	7	1
Male	Dublin	1	6	1
Male	Mid-East	1	7	1
Male	Mid-West	1	7	1
Male	Midlands	1	9	1
Male	South-East	1	8	1
Male	South-West	0	5	1
Male	West	0	6	1

As can be seen, there is a huge difference between genders in those who are looking after home/family. There are significant standard deviations present throughout, indicating the presence of outliers. It is apparent looking at the mean figures, that there are a greater number of Males who are Unemployed ('LFFJ' and 'ULGUP') than females throughout all regions. This was subsequently confirmed as below:

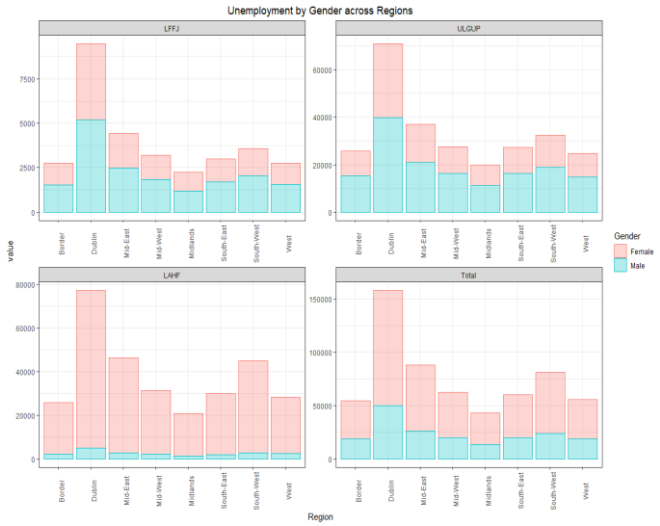
Gender	LFFJTot	ULGUPTot	LAHFTot	Total
Female	13957	112509	284809	411275
Male	17477	153453	20747	191677

Removing the Looking After Home & Family column results in the total number of persons unemployed by gender changing drastically:

Gender	LFFJTot	ULGUPTot	Total
<chr>	<int>	<int>	<dbl>
Female	13957	112509	126466
Male	17477	153453	170930

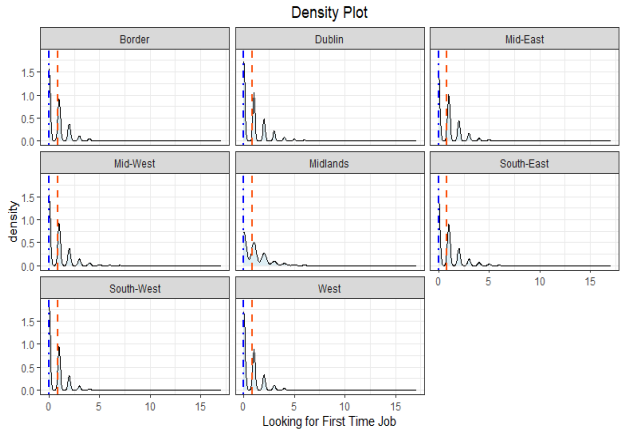
C. Visualisation

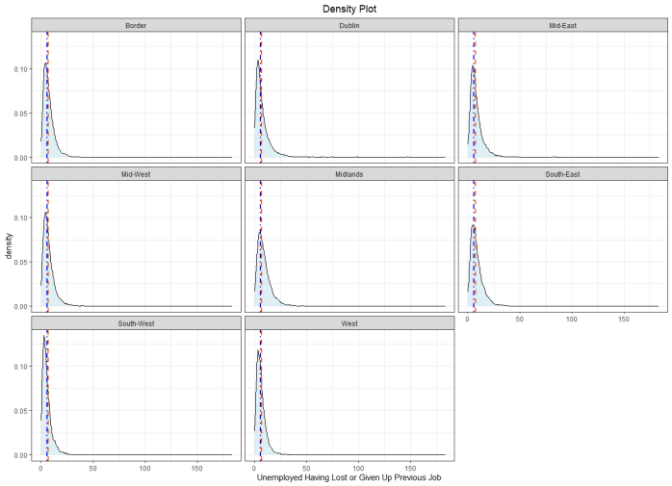
Initial visualisation began by melting the dataframe referenced earlier -'Insight2spruced', so as to enable an 'all in one' graph for the purpose of interpreting both insights at once, illustrating unemployment type by gender across regions:

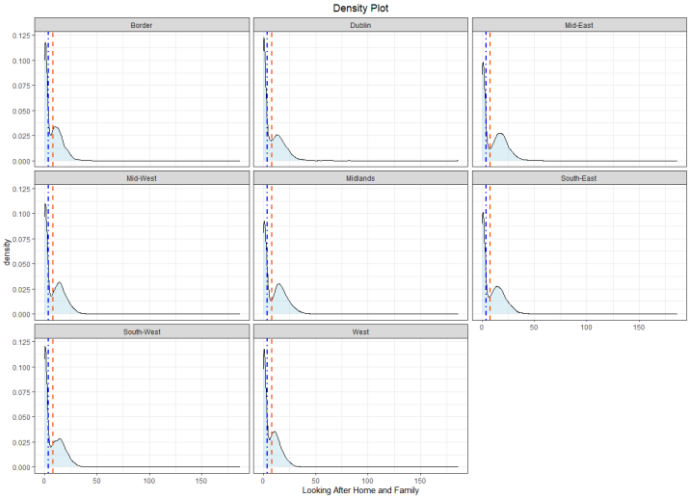


Given the enormous difference between genders across all regions for those who are looking after home/family as evidenced above; this type of unemployment was removed from analysis for unemployment by gender.

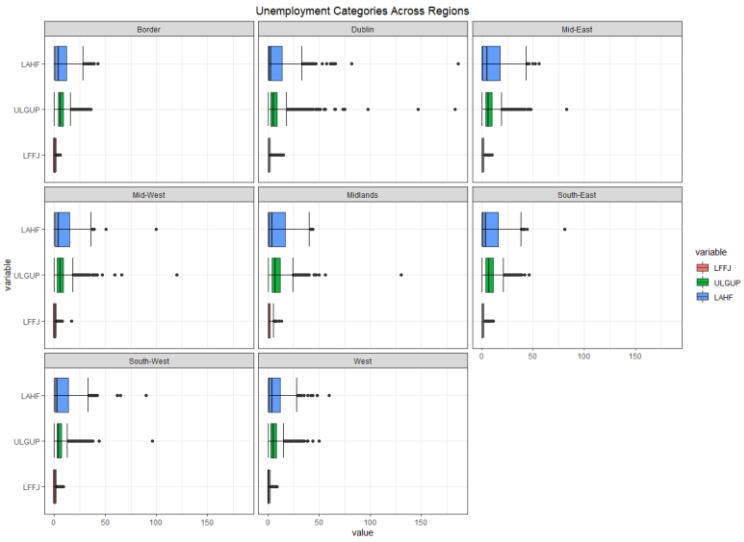
When illustrating the distribution of the data via density plots, it becomes apparent the data is positively skewed for LFFJ, LAHF and ULGUP, with a seemingly bimodal characteristic apparent for ULGUP, with the blue line representing the median, and the red line the mean:







Finally, a boxplot was created for those LFFJ, ULGUP and LAHF by region with error bars, thereby highlighting observations that fell outside of a 95% confidence interval:



D. Statistical Analysis of Unemployment by Region

Given the data gathered for unemployment by region appears to be highly skewed with high degrees of kurtosis, the Pearson Chi Test was used to confirm the data was non normally distributed (Thode, H. C. (2002). Through use of Sapply (to gather all results at once) the following results were generated where

H_0: The data is normally distributed

H_1: The data is non normally distributed

```
          LFFJ                                ULGUP
statistic 1758477                             319447.4
p.value   0                                   0
method    "Pearson chi-square normality test" "Pearson chi-square normality test"
data.name "X[[i]]"                            "X[[i]]"
n.classes 135                                 135
df        132                                 132
          LAHF
statistic 389928.5
p.value   0
method    "Pearson chi-square normality test"
data.name "X[[i]]"
n.classes 135
df        132
```

Results

As p < .05 for all distributions, we reject the null hypothesis and conclude data is non normally distrubuted.

Given these results, a non parametric test must be used. With this in mind, the Kruskal Wallis test (Wallis, 1952) was used to check if there are significant differences in each unemployment category by region where:

H_0: There is no difference between regions by unemployment status

H_1: There is a difference between regions by unemployment status

Results

<u>Looking for first time job by region</u>
H (7) = 438.76, P < 2.2e-16

<u>Unemployed having lost or given up job by region</u>
H (7) = 1195, P < 2.2e-16

<u>Looking after home and family by region</u>
H (7) = 186.89, P < 2.2e-16

As all three results have a p-value less than .05 we can determine significant differences are presrnt in each unemployment category, by region.

So as to identify where these differences exist, the Dwass-Steel-Critchlow-Fligner Test (Critchlow, D. E. and Fligner, M. A., 1991) was used. This test compares the median/means of all pairs of groups using the Steel-Dwass-Critchlow-Fligner pairwise ranking nonparametric method as below:

```
        Pairwise comparisons using Dwass-Steele-Critchlow-Fligner all-pairs test

data: LFFJ by Region

           Border  Dublin  Mid-East Mid-West Midlands South-East South-West
Dublin     2.3e-09 -       -        -        -        -          -
Mid-East   2.3e-09 0.99525 -        -        -        -          -
Mid-West   0.62047 0.00021 9.2e-05  -        -        -          -
Midlands   1.2e-12 0.05252 0.28544  5.3e-08  -        -          -
South-East 0.02093 0.24000 0.09715  0.75801  0.00024  -          -
South-West 9.0e-11 < 2e-16 < 2e-16  5.3e-14  < 2e-16  6.9e-14    -
West       0.01225 7.3e-14 8.1e-14  2.8e-06  7.1e-14  4.7e-10    0.02414

data: ULGUP by Region

           Border  Dublin  Mid-East Mid-West Midlands South-East South-West
Dublin     1.7e-07 -       -        -        -        -          -
Mid-East   1.8e-05 < 2e-16 -        -        -        -          -
Mid-West   0.85643 0.00049 3.3e-09  -        -        -          -
Midlands   6.9e-14 < 2e-16 6.8e-10  < 2e-16  -        -          -
South-East 4.6e-11 < 2e-16 0.13865  8.7e-14  0.00102  -          -
South-West < 2e-16 < 2e-16 < 2e-16  < 2e-16  < 2e-16  < 2e-16    -
West       9.0e-14 0.00083 < 2e-16  3.4e-13  < 2e-16  < 2e-16    6.5e-14

data: LAHF by Region

           Border  Dublin  Mid-East Mid-West Midlands South-East South-West
Dublin     1.00000 -       -        -        -        -          -
Mid-East   5.6e-14 6.6e-14 -        -        -        -          -
Mid-West   0.09304 0.05232 2.6e-09  -        -        -          -
Midlands   7.9e-10 3.9e-11 0.96575  0.00029  -        -          -
South-East 0.00022 2.0e-05 0.00132  0.48696  0.20749  -          -
South-West 0.99879 0.99774 8.1e-14  0.29206  2.3e-09  0.00054    -
West       0.99995 0.99597 6.7e-14  0.19811  4.4e-09  0.00088    1.00000
```

Pairs where p <.05:

i) Looking for first time job

```
                      Comparison  p.value
South-East - Border = 0   0.0209
      West - Border = 0   0.0123
  Mid-West - Dublin = 0 0.000214
 South-West - Dublin = 0        0
South-West - Mid-East = 0       0
South-East - Midlands = 0 0.000236
South-West - Midlands = 0       0
    West - South-West = 0   0.0241
```

ii) Unemployed having given up or lost previous job

```
                      Comparison  p.value
South-West - Border = 0        0
  Mid-East - Dublin = 0        0
  Mid-West - Dublin = 0 0.000491
  Midlands - Dublin = 0        0
South-East - Dublin = 0        0
South-West - Dublin = 0        0
      West - Dublin = 0 0.000828
South-West - Mid-East = 0      0
     West - Mid-East = 0       0
 Midlands - Mid-West = 0       0
South-West - Mid-West = 0      0
South-East - Midlands = 0  0.00102
South-West - Midlands = 0      0
     West - Midlands = 0       0
South-West - South-East = 0    0
     West - South-East = 0     0
```

iii) *Looking after Home and Family*

```
                     Comparison  p.value
South-East - Border = 0 0.000223
South-East - Mid-East = 0  0.00132
  Midlands - Mid-West = 0 0.000295
South-West - South-East = 0 0.000536
     West - South-East = 0 0.000883
```

E. Interpretation of Result

Significant differences in employment status by region exist throughout the selected data. Grouping by region, the least variance by way of p-values is the Looking after Home and Family employment status, followed by those looking for their first job and unemployed having given up or lost their previous job. Those looking for their first job in the South-West region stand out as being the greatest statistically significant grouping. Drawing from this, the inference can be made that those looking for their first-time job in the South-West stand out as being most disproportionately affected. As the South-West region includes Cork, this is a very interesting result as one would usually presume the presence of a major urban center would help propagate employment. Further to this, unemployment having lost or given up a job is most statistically significant in the Dublin region. As this is again, a region with a major urban center, it is notable considering the calculations chosen are not affected by sample size.

III. Unemployment By Gender

A. Statistical Analysis of Unemployment by Gender

Similar to before, the Kruskal Wallis test was used to check for significance in difference between genders who are LFFJ and ULGUP, generating the following results where:

H_0: There is no difference between genders by unemployment status

H_1: There is a difference between genders by employment status

Results

<u>Looking for first time job by Gender</u>
H (1) = 216.55, p < 2.2e-16

<u>Unemployed having lost or given up job by Gender</u>
H (7) = 1565.4, p < 2.2e-16

As p <.05 for both, we reject the null hypothesis and conclude the data is non normally distributed.

B. *Statistical Analysis of Unemployment by Gender by Region*

Building upon the analysis outlined in section II., I elected to compute the difference in unemployment by region **by gender.** In order to do this, the Wilcoxon Rank Sum Test (Wilcoxon, 1945). was used. In performing this test, I used two dataframes, one of which contained unemployment data by gender and region for Males, and the other Females. As the data was not paired, I specified the test was not to pair selected data. An example of the code used is as below:

```
wilcox.test(Insight2Females$LFFJ[which(Insight2Females$Region == "Dublin")], Insight2Males$LFFJ[which(Insight2Males$Region == "Dublin")], method = "bonferroni", paired = FALSE)
```

Hypothesis

H_0: There is no difference in unemployment status by gender

H_1: There is a difference in unemployment status by gender

Result

All regions had a statistically significant difference in unemployment by gender, with the exception of the Midlands for which those LFFJ had a p-value greater than or equal to .05.

```
        Wilcoxon rank sum test with continuity correction

data:  Insight2Females$LFFJ[which(Insight2Females$Region == "Midlands")] and Insight2Males$LFFJ[which(Insight2Mal
es$Region == "Midlands")]
W = 595030, p-value = 0.05331
alternative hypothesis: true location shift is not equal to 0
```

C. *Interpretation of Result*

In conclusion, there is a statistically significant difference in employment type by gender across all but one region (Midlands) and unemployment type (LFFJ). The summary statistics previously described lend credence to this result. Given 52% of Irelands citizens are females (CSO, 2016) it is striking that males are more disproportionately affected by unemployment as this variance cannot be accounted for as being due to Irelands gender composition. It could be a case that females who are identified as looking after their home and family have identified as this in place of being unemployed due to stigma and societal norms.

IV. Commute Leaving Time

A. *Pre-Processing For Both Commute Leaving Time and Commute Travel Time*

A similar approach to that previously taken in sections II. and III. was taken in preparing appropriate data for this analysis. Firstly, two new dataframes were created, one for commute travel time, and the other for commute leaving time with the columns renamed for ease of use and interpretation. Finally, both sets of data were combined using the cbind() function. The data was also melted to enable suitable visualization.

B. *Summary Statistics*

```
    Region b4_630_Mean T630_700_Mean T701_730_Mean T731_800_Mean T801_830_Mean T831_900_Mean T901_930_Mean
    Border     6.38295      6.579967      8.485335      17.07028      24.12618      32.54566      18.057001
    Dublin     9.871774     13.791889     18.960877     29.17329      40.62577      32.70320      7.972143
  Mid-East    14.492754     17.043052     18.842711     27.06607      36.77067      40.36914      14.372123
  Mid-West     7.286011      9.417057     12.432137     21.17629      30.65003      36.82943      12.873115
  Midlands    11.515233     10.364695     11.649642     20.78674      29.40412      36.65233      19.131720
South-East     9.193817      9.060642     11.425089     21.37099      31.72949      35.48751      13.966706
South-West     7.262028     10.261336     13.940464     22.59709      32.26134      33.42748      10.725857
      West     6.638277      8.599198     11.136774     19.19038      25.24198      33.63377      15.729459
T930aft_Mean
    8.982291
   14.706063
   12.050298
    9.999480
    9.796595
    9.381688
   10.271374
   11.666834
```

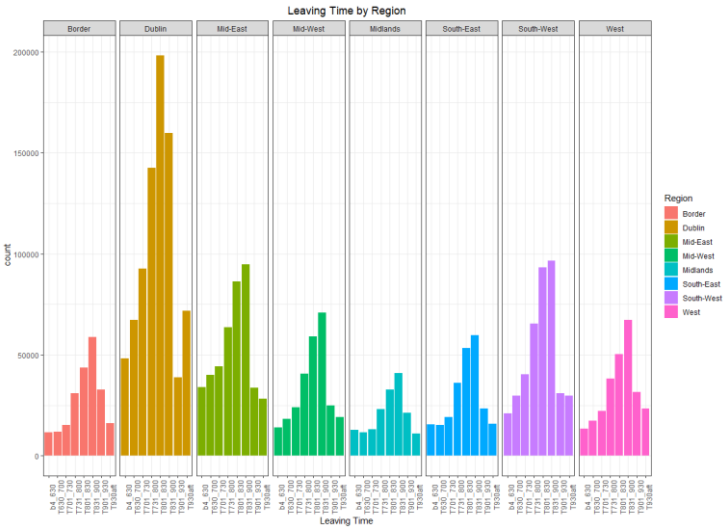
```
    Region b4_630_Median T630_700_Median T701_730_Median T731_800_Median T801_830_Median T831_900_Median
    Border             5              6.0              8             16             21             30
    Dublin             8             13.0             18             27             37             30
  Mid-East            13             16.0             18             26             34             37
  Mid-West             7              9.0             12             19             27             34
  Midlands            10              9.0             11             20             27             34
South-East             8              8.5             11             20             28             32
South-West             6              9.0             13             20             28             29
      West             6              8.0             10             16             21             31
T901_930_Median T930aft_Median
             16              8
              7             13
             11             11
             10              9
             17              9
             11              8
              8              9
             13             10
```

```
    Region b4_630_SD T630_700_SD T701_730_SD T731_800_SD T801_830_SD T831_900_SD T901_930_SD T930aft_SD
    Border  5.629099    4.274120    4.866510    9.141743    13.41738    18.43624    11.021049  5.412622
    Dublin  7.050985    7.474968    9.353224   13.945994    18.69617    15.99327     5.437888  7.785819
  Mid-East  7.286474    8.108346    9.442610   12.689951    18.06747    20.10024    11.465125  6.913109
  Mid-West  4.886149    5.237716    6.789150   12.412844    18.21997    19.39230     9.829426  6.248621
  Midlands  7.218359    5.671467    6.161650   10.264714    16.32920    17.27745    12.229481  5.288012
South-East  5.503058    4.818034    6.023406   10.781329    18.36425    19.82371    10.359801  5.516023
South-West  4.846631    6.103780    8.384992   13.963930    19.57564    18.88747     8.634645  6.890850
      West  4.361993    5.196748    7.128244   12.519051    16.65312    17.98832    10.828578  8.216904
```

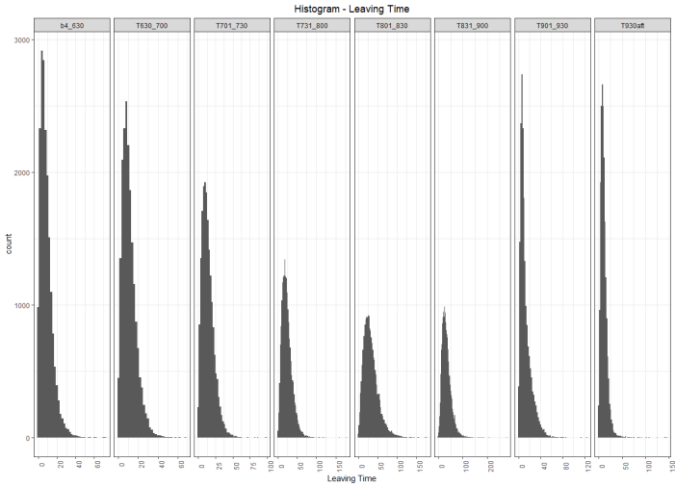
From review of the above descriptive statistics, it would seem a greater number of people travel early in Dublin, the Mid-East, Midlands and South-East when reviewing the mean for each time interval, however this would be naive to conclude. It it is only natural areas of high density such as Dublin appear this way, when compared with regions such as the South-West and the West as these parameters are not normalized by population size per region. While one could do this, it is unnecessary as one can just run a statistical test that is not affected by sample size which would have the same effect. Further to this, given the standard deviation across the majority of time intervals, it suggests large outliers are present which could otherwise skew our analysis.

C. *Visualisation*

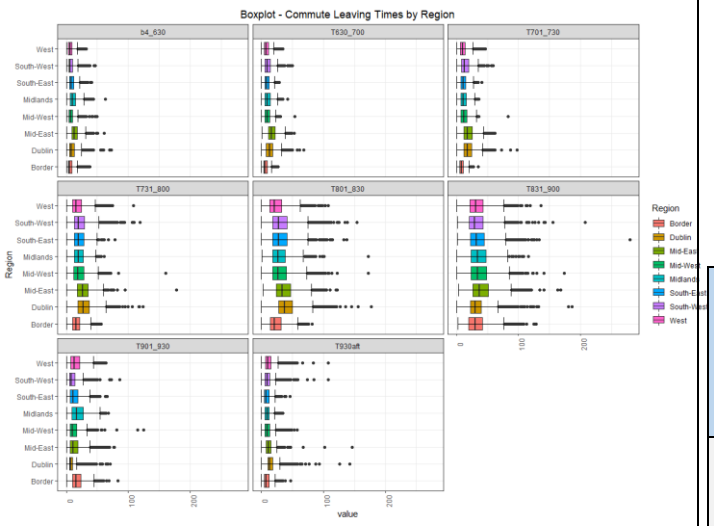
The below bar chart illustrates the trends in commute leaving time across regions. A significant proportion of people commute to work between 07:30 and 09:00 throughout all regions.



In place of a density plot, a histogram was generated to illustrate the distribution of data for each time interval. The data is positively skewed for all regions, with significant levels of kurtosis present. This suggests the need to use a non-parametric test when examining differences between regions by way of commute time.



Given the apparent presence of outliers from analysis of the standard deviation between time intervals by way of regions, a boxplot was generated with a 95% confidence interval and Inter-Quartile Range (IQR) as below. A significant number of outliers are present throughout the various time intervals, which needs to be accounted for in the methodology chosen for testing between groups.



D. Statistical Analysis of Leaving Time by Region

To confirm the distribution of data is non normal, the Anderson Darling Test (Anderson, T. W.; Darling, D. A. , 1952) utilized giving the following results where;

H_0: The data is normally distributed

H_1: The data is non normally distributed

```
          b4_630                              T630_700
statistic 432.8314                            295.9494
p.value   3.7e-24                            3.7e-24
method    "Anderson-Darling normality test"   "Anderson-Darling normality test"
data.name "X[[i]]"                            "X[[i]]"
          T701_730                            T731_800
statistic 257.7603                            231.5198
p.value   3.7e-24                            3.7e-24
method    "Anderson-Darling normality test"   "Anderson-Darling normality test"
data.name "X[[i]]"                            "X[[i]]"
          T801_830                            T831_900
statistic 250.0299                            266.8623
p.value   3.7e-24                            3.7e-24
method    "Anderson-Darling normality test"   "Anderson-Darling normality test"
data.name "X[[i]]"                            "X[[i]]"
          T901_930                            T930aft
statistic 816.697                             389.7763
p.value   3.7e-24                            3.7e-24
method    "Anderson-Darling normality test"   "Anderson-Darling normality test"
data.name "X[[i]]"                            "X[[i]]"
```

As P < .05 for all time intervals, the data being analysed is not normally distributed.

Given this, and the presence of a significant number of outliers, Moods Median Test (Mood, A.M., 1954) was used. This test is not as affected by outliers as Kruskal Wallis given the Moods Median Test is based on the Median, while the Kruskal Wallis uses the Mean which renders it is more exposed to outliers.

H_0: There are no differences in leaving times by region

H_1: There are differences in leaving times by region

```
        Mood's median test

data:  b4_630 by Region
X-squared = 2073.3, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T630_700 by Region
X-squared = 2455.3, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T701_730 by Region
X-squared = 2546.9, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T731_800 by Region
X-squared = 1563.9, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T801_830 by Region
X-squared = 1497.2, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T831_900 by Region
X-squared = 268.21, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T901_930 by Region
X-squared = 2221.6, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T930aft by Region
X-squared = 1564.3, df = 7, p-value < 2.2e-16
```

As p < 0.5, there are statistically significant differences between regions across all time interval.

Using the Dunn's Test (Dunn, O.J ,1964) post-hoc to determine where differences lies between groups revealed that statistically significant differences exist amongst all regions by time interval**. Non significant** results are noted on the following table (where p => .05):

Travel Interval	Regions
Before 06:30	Mid West ~ South West
06:31 - 07:00	Mid West ~ South East Midlands ~ South West
07:01 - 07:30	Dublin ~ Mid East Mid West ~ Midlands Midlands ~ South East South East ~ West
07:31 - 08:00	Mid West ~ Midlands Mid West ~ South East Mid West ~ South West Midlands ~ South East Midlands ~ South West South West ~ South East

08:01 - 08:30	Border ~ West Mid West ~ Midlands Mid West ~ South East Mid West ~ South West Midlands ~ South East South Eadt ~ South West
08:31 – 09:00	Border ~ Dublin Border ~ South West Dublin ~ South West Dublin ~ West Mid West ~ Midlands Mid West ~ South East Midlands ~ South East South East ~ West
09:01 – 09:30	Border ~ Midlands Mid East ~ South East Mid East ~ Mid West Mid West ~ South East
After 09:30	Border ~ South East Mid West ~ Midlands Mid West ~ South West Midlands ~ South East Midlands ~ South West

E. Interpretation of Results

As illustrated in the above table, the greatest variation in travel times exists outside of the 07:30 – 09:00 time period. This reflects our earlier observation which showed greater numbers of people travel between 07:30 – 09:00 meaning there is a greater level of homogeneity during this time period, with similar inflection points.

V. Commute Travel Time

A. Summary Statistics

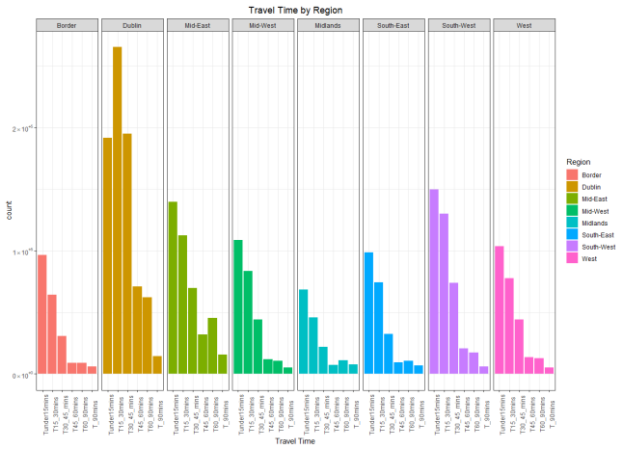
Region	Tunder15mins_Mean	T15_30mins_Mean	T30_45_mins_Mean	T45_60mins_Mean	T60_90mins_Mean	T_90mins_Mean
<chr>	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*
Border	53.4	35.6	16.9	4.85	4.90	3.23
Dublin	39.2	54.3	39.9	14.5	12.8	2.98
Mid-East	59.5	48.0	29.8	13.6	19.3	6.73
Mid-West	56.4	43.5	22.9	6.28	5.61	2.60
Midlands	61.2	41.1	19.7	6.66	10.0	6.78
South-East	58.7	44.2	19.4	5.46	6.35	4.17
South-West	51.8	45.1	25.6	7.16	5.94	2.06
West	51.9	39.0	22.0	6.71	6.46	2.68

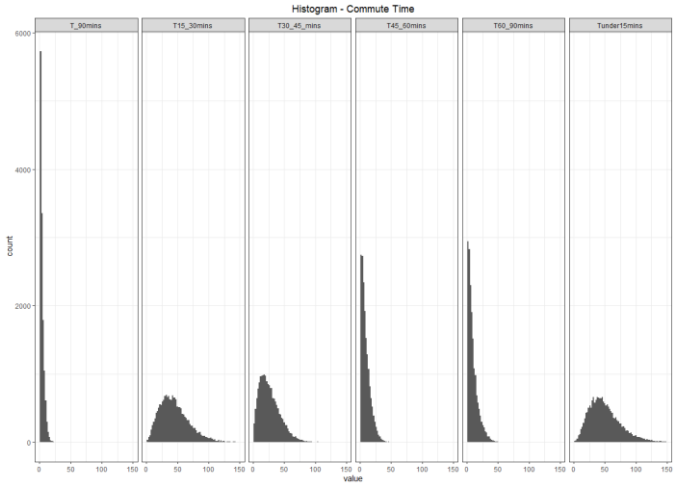
Region	Tunder15mins_Me~	T15_30mins_Medi~	T30_45_mins_Med~	T45_60mins_Medi~	T60_90mins_Medi~	T_90mins_Median
<chr>	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*
Border	49	32	15	4	4	2
Dublin	36	51	38	13	11	2
Mid-Ea~	55	45	28	12	18	6
Mid-We~	52	38	20	5	5	2
Midlan~	58	39	18	6	8	6
South-~	53	40	17	5	5	4
South-~	48	41	22	6	5	2
West	49	33	18	5	6	2

Region	Tunder15mins_SD	T15_30mins_SD	T30_45_mins_SD	T45_60mins_SD	T60_90mins_SD	T_90mins_SD
<chr>	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*
Border	27.3	20.1	9.94	3.48	4.03	2.96
Dublin	21.6	22.0	16.4	7.98	8.28	2.76
Mid-East	26.2	21.4	14.9	8.23	10.1	3.86
Mid-West	28.4	25.9	14.7	4.72	4.17	2.28
Midlands	26.3	19.1	10.3	4.67	7.44	4.10
South-East	28.9	24.7	11.8	4.16	5.14	3.17
South-West	26.1	26.4	16.5	6.01	5.10	1.82
West	25.1	23.9	15.8	5.76	4.58	2.16

Standard Deviation for each commute interval gets larger as the length of time for the commute gets smaller. This is also echoed in both the Median and Mean across commute intervals. This indicates that the data is heavily skewed and as such nonparametric testing may be required. The large standard deviation when compared to mean values indicates as in the commute leaving time analysis, that there is potentially a significant number of outliers.

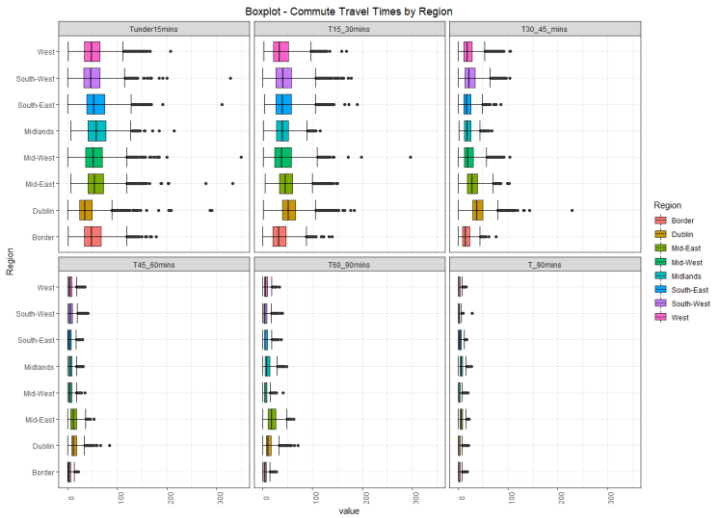
B. Visualisation





The above two plots highlight the distribution of values between each commute interval, and also mirrors the observation noted from summary statistics. The degree of skewness becomes greater the shorter one's commute takes.

The below boxplot generated in the same vein as before (containing a 95% confidence interval and standard error bars) shows that there are a significant number of outliers present throughout this data.



C. Statistical Analysis of Commute Travel Times by Region

As before, the Anderson Darling test was used to confirm the above observations where:

H_0: The data is normally distributed

H_1: The data is non normally distributed

```
          Tunder15mins                       T15_30mins
statistic 232.4825                           170.2033
p.value   3.7e-24                           3.7e-24
method    "Anderson-Darling normality test"  "Anderson-Darling normality test"
data.name "X[[i]]"                           "X[[i]]"
          T30_45_mins                        T45_60mins
statistic 244.7223                           509.756
p.value   3.7e-24                           3.7e-24
method    "Anderson-Darling normality test"  "Anderson-Darling normality test"
data.name "X[[i]]"                           "X[[i]]"
          T60_90mins                         T_90mins
statistic 713.1297                           683.7036
p.value   3.7e-24                           3.7e-24
method    "Anderson-Darling normality test"  "Anderson-Darling normality test"
data.name "X[[i]]"                           "X[[i]]"
```

As $p < .05$, the data is non normally distributed.

Given these results, and the presence of significant outliers, I elected to use the Moods Median Test once more so as to determine if there are significant differences in commute travel times by region generating the following results:

```
        Mood's median test

data:  Tunder15mins by Region
X-squared = 1334.1, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T15_30mins by Region
X-squared = 1105.7, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T30_45_mins by Region
X-squared = 4006.8, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T45_60mins by Region
X-squared = 4293.7, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T60_90mins by Region
X-squared = 4191.9, df = 7, p-value < 2.2e-16

        Mood's median test

data:  T_90mins by Region
X-squared = 3177.8, df = 7, p-value < 2.2e-16
```

Given these results, the Dunn's Test was utilised again to identify where differences between regions by time interval occur. **Non-significant results** are captured in the following table (where p =>.05):

Commute Time	Regions
Under 15 minutes	Border ~ South West Border ~ West Mid-East ~ Midlands Mid-East ~ South East Mid-West ~ South East
Between 15 and 30 minutes	Mid-West ~ Midlands Mid-West ~ South-East Mid-West ~ South-West Midlands ~ South-East Midlands ~ South-West South-East ~ South-West
Between 30 and 45 minutes	Mid-West ~ West Midlands ~ South-East Midlands ~ West
Between 45 and 60 minutes	Mid-West ~ Midlands Mid-West ~ West Midlands ~ South-East Midlands ~ West
Between 60 and 90 minutes	Mid-West ~ South-East Mid-West ~ South-West South-East ~ South-West South-East ~ West
Greater than 90 minutes	Dublin ~ West Mid-East ~ Midlands Mid-West ~ West

D. Interpretation of Results

When Dublin and the Border regions are compared against other regions in the context of the above results, it becomes apparent that they are the regions with the greatest statistically significant difference from others, that is – they are notable, the 'outliers' amongst regions in that almost all of their commute times are different across all time intervals with the Border region having a statistically significant difference across all results with one exception, and Dublin much the same save for two exceptions.

While traffic congestion could account for these variances in the Dublin region, this does not apply, or at least not nearly as affectual in the context of the Border region.

VI. Home Ownership

A. Pre-Processing

In order to compare home ownership levels across regions, the following code was used:

```
Insight5 <- Df[,c(392,393,399, 809,811)]

Insight5$Homeowners <- rowSums(subset(Insight5, select = c("T6_3_OMLH", "T6_3_OOH")))

Insight5$NotHomeowners <- (Insight5$T6_3_TH - Insight5$Homeowners)

#Ratio of Homeowners to Non Homeowners

Insight5$ratio <- Insight5$Homeowners/(Insight5$NotHomeowners + Insight5$Homeowners)
```

In order to enable the analysis I had in mind (ratio analysis), a ratio was calculated between those who owned a home, and those who did not. This was an important step to take, as it reduced the effort that would otherwise be needed to transform the selected data in its absence.

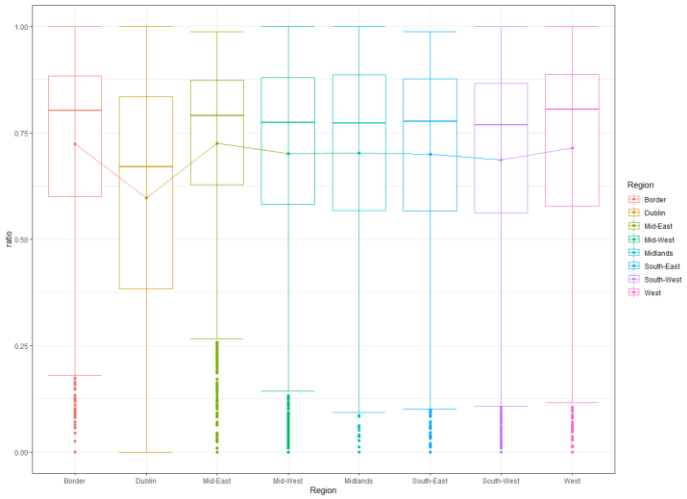
B. Summary Statistics

Region	count	mean	median	sd	IQR
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Border	1807	0.725	0.803	0.204	0.283
Dublin	4882	0.598	0.671	0.277	0.452
Mid-East	2346	0.725	0.790	0.199	0.246
Mid-West	1923	0.702	0.775	0.226	0.297
Midlands	1116	0.703	0.773	0.223	0.318
South-East	1682	0.700	0.777	0.223	0.311
South-West	2889	0.687	0.768	0.232	0.303
West	1996	0.714	0.806	0.224	0.309

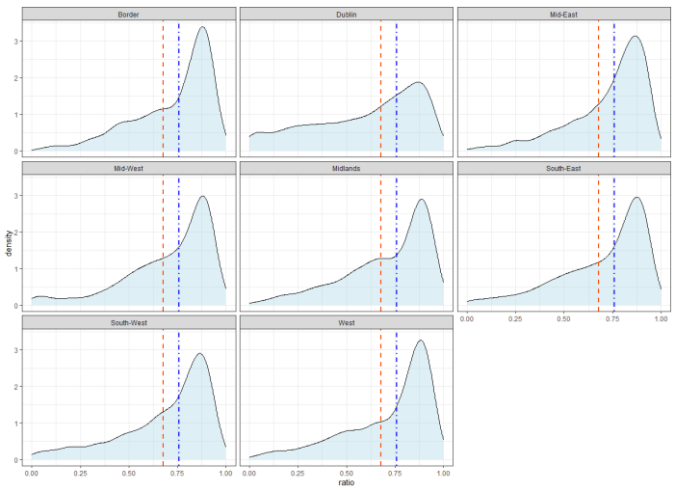
From review of these statistics, it is apparent that home ownership levels in Dublin are significantly less than other regions. The mean, median and standard deviations remain relatively similar across all other regions. The IQR is far greater in Dublin than all other regions indicating that home ownership levels are more variable in Dublin than other regions.

C. Visualisation

In the below boxplot, we can see that outliers are present in all regions with the exception of Dublin.



As the boxplots indicate, the presence of outliers is skewing our data as illustrated in the below density graph where blue represents the median, and red the mean:



D. Statistical Analysis of Home Ownership Levels by Region

Given the values for the ratio appear heavily skewed throughout all regions, the Cramer Von Mises test (Cramer; Von Mises, 1928) for normality was used where;

H_0: The data is normally distributed
H_1: The data is non normally distributed

providing the following result:

```
        Cramer-von Mises normality test

data:  Insight5$ratio
W = 113.63, p-value = 7.37e-10
```

As the normality assumption was violated ($p < .05$) , the Kruskal test was used to determine if there are statistically signifiance differences in home ownership levels throughout regions where:

H_0: There is no difference between ownership levels by region
H_1: There is a difference between ownership levels by region

```
        Kruskal-Wallis rank sum test

data:  ratio by Region
Kruskal-Wallis chi-squared = 601.73, df = 7, p-value < 2.2e-16
```

To identify where these statistically significant differences arise, the Conover-Iman Test (Conover-Iman,1979) of Multiple Comparisons Using Rank Sums was used. This is a particularly appropriate post- hoc test for Kruskal Wallis, as it is built on the same type of analysis (Rank Sums). Calling this test with the Bonferroni adjustment method selected, resulted in the following:

Col Mean- Row Mean	Border	Dublin	Mid-East	Mid-West	Midlands	South-Ea
Dublin	16.99553 0.0000*					
Mid-East	0.678249 1.0000	-17.78403 0.0000*				
Mid-West	2.650159 0.1127	-14.15747 0.0000*	2.132484 0.4617			
Midlands	1.821059 0.9606	-12.01518 0.0000*	1.322807 1.0000	-0.464954 1.0000		
South-Ea	2.957846 0.0434	-13.00800 0.0000*	2.472205 0.1881	0.401010 1.0000	0.799942 1.0000	
South-We	5.328245 0.0000*	-13.12939 0.0000*	4.986245 0.0000*	2.479705 0.1842	2.567068 0.1437	1.942969 0.7285
West	0.312864 1.0000	-17.23272 0.0000*	-0.363524 1.0000	-2.399377 0.2301	-1.583095 1.0000	-2.720823 0.0913

E. Interpretation of Results

As can be seen above, Dublin stands out amongst all other regions with there being a statistically significant difference in home ownership levels between it and all other regions. Two other differences in home ownership levels are apparent, those between the Mid-East and South-West and finally the Border and South-West regions.

It is worthwhile noting that while home ownership levels in Dublin are significantly different from all other regions, if one were to exclude Dublin from analysis – only two of 21 pairs would be statistically significant indicating that with the exception of Dublin, home ownership levels on the whole are homogenous across Ireland. In Dublin, the supply of housing is not meeting demand, and there are a greater number of students present than across other regions which combined, could account for this drastic difference.

VII. Conclusion

In conclusion, those who are unemployment having lost or given up a job is most statistically significant in the Dublin region while looking for their first job in the South-West region stand out as being the greatest statistically significant grouping.

When it comes to Gender, there are significant differences amongst all regions with the exception of the Midlands.

Regarding commute time, most people across all regions commute between 07:30 and 09:00, with more homogeneity than those who commute outside of this time period. For travel time, the Border and Dublin regions have the greatest disparity in travel times when paired with other regions.

Home ownership levels are mostly homogenous across Ireland, with Dublin being an outlier in this regard.

Finally, there is plenty of scope for further analysis on the selected insights, for example:

- Multilinear Regression between travel time, and commute time by engaging in log transformation
- Removing rows where student size is above a certain parameter and reducing the non-homeowner column rows by the extracted rows (row-wise)
- Statistical analysis of job category by gender and it's relation to the results of employment category by gender

VIII. References

Anderson, T. W.; Darling, D. A. (1952). "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes". Annals of Mathematical Statistics

Conover, W. J. and Iman, R. L. (1979) On multiple-comparisons procedures. Technical Report LA-7677-MS, Los Alamos Scientific Laboratory

Critchlow, D. E. and Fligner, M. A., 1991: On distribution-free multiple comparisons in the oneway analysis of variance. Communications in Statistics – Theory and Methods 20, 127–139

Cramér, H. (1928). "On the Composition of Elementary Errors". Scandinavian Actuarial Journal. 1928

CSO. (2016). Society - CSO - Central Statistics Office. [online] Available at: https://www.cso.ie/en/releasesandpublications/ep/p-wamii/womenandmeninireland2016/society/#d.en.139033

Dunn, O. J. 1964. Multiple comparisons using rank sums. Technometrics 6: 241–252.

Kruskal, Wallis (1952). "Use of ranks in one-criterion variance analysis". Journal of the American Statistical Association, (260): 583–621

Mood, A. M. (1954). On the asymptotic efficiency of certain nonparametric two-sample
tests. The Annals of Mathematical Statistics, 25(3):514–522.

Thode, H. C. (2002). Statistics: textbooks and monographs, Vol. 164.Testing for normality.

Von Mises, R. E. (1928). Wahrscheinlichkeit, Statistik und Wahrheit

Wickham, H., 2014. Tidy Data. Journal of Statistical Software, [online] Available at: https://vita.had.co.nz/papers/tidy-data.pdf

Wilcoxon, Frank (1945). "Individual comparisons by ranking methods"

IX. Appendix

1. Kruskal Wallis test for unemployment status by region

```
        Kruskal-Wallis rank sum test

data:  LFFJ by Region
Kruskal-Wallis chi-squared = 438.76, df = 7, p-value < 2.2e-16

        Kruskal-Wallis rank sum test

data:  LAHF by Region
Kruskal-Wallis chi-squared = 186.89, df = 7, p-value < 2.2e-16
```

2. Kruskal Wallis test for gender by unemployment status

```
        Kruskal-Wallis rank sum test

data:  LFFJ by Gender
Kruskal-Wallis chi-squared = 216.55, df = 1, p-value < 2.2e-16

        Kruskal-Wallis rank sum test

data:  ULGUP by Gender
Kruskal-Wallis chi-squared = 1565.4, df = 1, p-value < 2.2e-16
```