Statistical Analysis of Survey on Attitude Towards Innovative Transport in Europe

1st Saumya Gautam,
MCM1, School of Computing,
Dublin City Univeristy
Dublin, Ireland
saumya.gautam3@mail.dcu.ie
Student ID: 19211275

2nd Saumitra Das
MCM1, School of Computing
Dublin City University
Dublin, Ireland
Saumitra.das2@mail.dcu.ie
Student ID: 19211286

***Abstract*— EU Transport data from a survey done on 28 European countries is used to deduce the relationship between country regions and their affiliation with the possibility of buying an electric car in future, inclination towards car-sharing and consequently their relative concern towards the impact of cars on the environment. Chi-squared value and p-value are calculated to conclude that the country region has an impact on the outlook of individuals on above-said variables.**

Keywords— EU, Transport, Car-sharing, Electric car, Environment Concern, Chi-squared test

I. INTRODUCTION

Public transport is one of the backbones of urban mobility across the 28 member countries of the EU over relatively short distances, mainly in urban and suburban areas. EU transport policy aims to provide European mobility solutions that are reliable, secure and environmentally friendly. Congestion of road transport, development, commuter rights and infrastructure funding are just a few manifestations of EU-level transport issues. Projects such as ETIS and ETIS plus have gathered a plethora of transport data and useful information required to measure progress towards these goals. Along with the need to study trends in the transport system, the environment is also a significant concern amongst the general public and government bodies in the EU.

TRT Trasporti e Territorio and Ipsos conducted an EU wide transport survey with all 28 European countries in June 2014 to collect data on car use, the use of modes of transport for long-distance mobility as well as some other policy issues of concern under the supervision of the Joint Research Center – Institute for Perspective and Technological Studies (JRC-IPLS) of the European Commission. The survey also included several other aspects, such as their concern for the environment, the stance on internalizing external road costs by road charges as well as on innovative transport, including electric or hybrid cars and car-sharing.

The analysis of this paper is based on the innovative transport aspect of the study mentioned above. It will cover the study of responses of participants about their take on car-sharing, willingness to buy electric cars and their concern for the environment.

II. RELATED WORK

In June 2014, TRT Trasporti e Territorio and Ipsos conducted an EU Travel Survey on demand for the innovative transport system, with all 28 European countries to collect data on car use, on use of transport modes for long-distance mobility as well as on some other policy-relevant issues [1]. Concerning this paper's study, the survey findings were: The propensity of buying electric cars is generally higher in South Europe countries and often lower in North Europe. The propensity for car sharing is higher among those who use a car in combination with public transport and train.

"LESS CARS IS MORE" [2] this paper presents research and gathers data, evidence, and trends of the four revolutions of transport to model their effect on the activity at a vehicle (or passenger) level in several scenarios. It draws conclusions on climate changes concerning cars, innovation in electric car and car-sharing in Europe.

Barbora Bondorová & Greg Archer wrote a paper on "Does sharing cars really reduce car use?"[3] and the evidence shows ride-sharing apps do reduce the numbers of vehicles on the road and vehicle kilometres driven. Long-distance car-sharing services do compete with rail and coach services they also significantly increase car occupancy and reduce emissions per kilometer.

III. DATASET AND EXPLORATORY ANALYSIS

3.1 Dataset Information

The dataset was taken from EU Open data Portal on the study "EU Travel Survey on demand for innovative transport systems". It is an EU-wide survey carried out in 2014 with the objective of gathering several transport and mobility indicators on transport user preferences with emphasis on the potential of emerging transport technologies and the acceptability of various transport policy measures. The data used in the analysis are the elaboration of responses from a sample of 1000 individuals in each country (500 individuals in Cyprus, Luxembourg and Malta). The sample was segmented according to specific socio-economic characteristics in each region. In order to increase the reliability of the estimates, a stratified sample was described rather than a simple random sample by applying a weighting procedure to ensure that the responses were estimated on a sample reflecting the composition of EU adult population (from 16 years on) in terms of gender, age class, employment status and living region [1]. Therefore, the quality of data is excellent and reliable for statistical data analysis. The dataset is mainly categorical so we will be using contingency tables, multiple regression analysis and chi-squared test to analyze the dataset.

3.2 Exploratory Data Analysis

Exploratory data analysis has presented some relationship between the European countries and their opinions for considering the electric or hybrid car in their next purchase and for subscribing to car-sharing option. The result of this

relationship further showed some indication towards the European countries' concern on environmental impacts.

3.2.1 Feature Engineering

Application of feature engineering based on geographic domain knowledge is applied to categorize the European countries into four regions (East, West, North and South) for acquiring significant results.

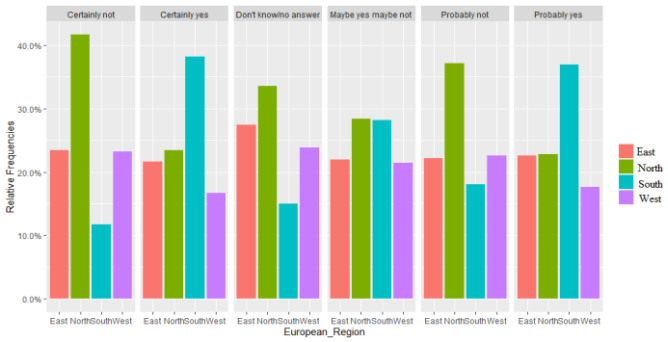
3.2.2 Summary Results of Plots and Contingency tables

3.2.2.1 European regions and their responses for considering electric/hybrid cars

Table1 and Figure1 depict that the Southern region has shown more intent on buying the electric/hybrid cars than other regions with 29.90 % response for "Probably yes" and 15.13 % response for "Certainly yes" (the highest percentage among the four regions). Whereas, Northern region has shown least interest for the same (16.13 % response for "Probably yes" and 8.13 % response for "Certainly yes"). Similarly, the Southern region's response towards "Probably not" and "Certainly not" is the least among other regions while the Northern region's response is highest for not considering electric/hybrid cars. East and West European countries have provided moderate responses to all the categories for considering electric/hybrid car. Although, Eastern region have shown more inclination towards considering electric/hybrid cars than western region.

Consider Electric car	East	North	South	West
Certainly not	10.26	13.74	4.41	11.13
Certainly yes	9.95	8.13	15.13	8.39
Don't know/no answer	9.73	8.94	4.58	9.26
Maybe yes maybe not	29.25	28.41	32.29	31.28
Probably not	19.5	24.61	13.67	21.73
Probably yes	21.27	16.13	29.9	18.17

<Table1: Proportion table of European Regions considering electric or hybrid vehicle on next purchase >



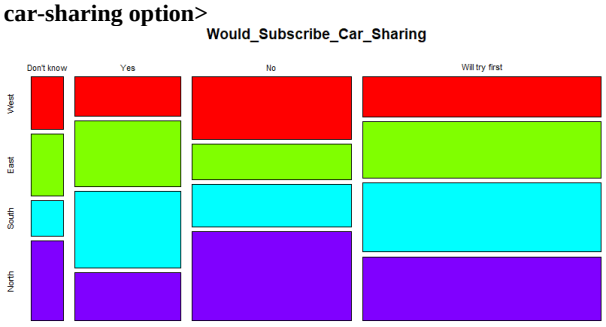
<Figure1: Relative Frequencies of European Regions for their opinions of considering Electric/Hybrid car>

3.2.2.2 European regions and their responses for considering car-sharing option

Contingency Table2 and Figure2 show that Southern and Eastern region depict inclination towards car sharing with 24.99% and 24.90% response to the "Yes" category respectively. These two regions have also provided less percentage of response towards the "No" category. In contrast, the Northern and Western region have shown least interest in the car-sharing option with a lesser percentage of responses for the "Yes" category and a higher percentage of response for the "No" category. Moreover, the highest frequency of people has given their opinion to try out the car-sharing option first and then decide.

	East	North	South	West
Don't know	7.03	6.89	3.49	6.63
No	20.03	38.29	20.77	39.30
Try First	48.02	41.07	50.73	37.65
Yes	24.90	13.73	24.99	16.39

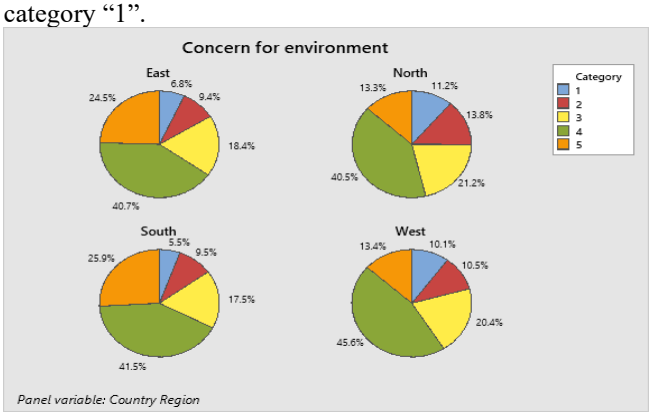
<Table2: Proportion table of European Regions subscribing to car-sharing option>



<Figure2: Mosaic plot showing frequencies of European Regions for their opinion of subscribing to car-sharing option >

3.2.2.3 European regions and their concern rating towards Environmental impacts

European people have provided rating to their level of concern for the impact of traffic on the environment. On a scale from 1 (not concerned) to 5 (very concerned), the average rate is very close to 4. Variability across countries is limited; the average is always between 3 and 4. However, Figure3 depicts that Southern and Eastern have shown more significant concern on environmental impact than Northern and Western region as the former gave high response to the category "5", whereas the latter gave high response to the category "1".



<Figure3: European regions rating the level of concern towards environmental impacts>

3.2.2.4 Regression Analysis

Multiple Linear regression is the most widely used statistical technique. It is a way to model relationship between sets of variables. A multiple regression model is built to establish the relationship between categorical and one response (dependent variable); additional steps are needed to ensure

that the results are interpretable. The categorical data was recoded to numerical data to be used in the analysis. The regression equation is as follows:

Regression Equation

Concern_environmental_impacts = (3.5417 + 0.0 Would_subsribe_car_sharing_if_a_Don't KNow- 0.0056 Would_subsribe_car_sharing_if_a_No+ 0.2014 Would_subsribe_car_sharing_if_a_Will try first+ 0.3385 Would_subsribe_car_sharing_if_a_Yes+ 0.0 Know_what_car_sharing_is_No- 0.108 Know_what_car_sharing_is_Unsure/ no answer+ 0.0916 Know_what_car_sharing_is_Yes+ 0.0 Country Region_East - 0.2817 Country Region_North+ 0.0105 Country Region_South - 0.2079 Country Region_West- 0.07169 Number_vehicles_in_household+ 0.0 Considering_electric_or_hybrid__Confused- 0.2012 Considering_electric_or_hybrid__No+ 0.2265 Considering_electric_or_hybrid__Yes)

The Multiple Regression Model

Multiple regression is a linear transformation of the X variables such that the sum of squared deviations of the observed and predicted Y is minimized. The prediction of Y is accomplished by the following equation:

Y'i = b0 + b1X1i + b2X2i + ... + bkXki

The "b" values are called regression weights and are computed in a way that minimizes the sum of squared deviations.

$$\sum_{i=1}^{N}(Y_i - Y'_i)^2$$

Term	Coefficients	P-Value
Constant	**3.5417**	**0**
Would subscribe car sharing		
No	**-0.0056**	**0.858**
Will try first	**0.2014**	**0**
Yes	**0.3385**	**0**
Know what car sharing is		
Unsure/ no answer	**-0.108**	**0.479**
Yes	**0.0916**	**0**
Country Region		
North	**-0.2817**	**0**
South	**0.0105**	**0.611**
West	**-0.2079**	**0**
Number vehicles in household	**-0.07169**	**0**
Considering electric or hybrid		
No	**-0.2012**	**0**
Yes	**0.2265**	**0**

<Table 3 Coefficient and P-value for categorical and continuous predictors>

For this, R^2 (R-sq)= 6.91%, which means that the independent variables, explains 6.91% of the variability of the dependent variable. Adjusted R^2 is also an estimate of the effect size, which at 6.88%, is indicative that even high-variability data can have a significant trend. In this example, the regression model is statistically significant, $p < .05$ for most of the independent variables. When a predictor has a low p-value, it is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

In the values shown in <Table 3> , the predictor variables of Concern environmental impacts are significant because of their p-values are 0.000 and less than significance level.

R^2 is calculated by the using the formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Key output:

As per the multiple regression analysis output, when R-squared is low, low P values still indicate a real relationship between the significant predictors and the response variable.

IV. HYPOTHESE AND RESEARCH QUESTIONS

Research question: If there is a relationship between European Regions and, their opinion on considering an electric car / hybrid car and subscribing to the car-sharing option, then does that relationship show any propensity towards their level of concern on environmental impact? For answering the research question, three hypothesis testing are carried out.

Confidence Interval and P-value for the three Hypothesis tests

Since the full sample is used (1000 or 500 individuals), the interval of confidence around estimates is no more than 2 percent or 3 percent in the 95 percent of cases regardless of the actual value of the population size [1]. Therefore, a significance level of 5% is considered for all the hypothesis tests.

Method description for the three Hypothesis tests

In the form of the chi-squared distribution, the test statistics are interpreted with the required number of degrees of freedom as follows:

- If Statistic > Critical Value: significant result, reject null hypothesis (H0), dependent.
- If Statistic <= Critical Value: not significant result, fail to reject null hypothesis (H0), independent.

For dealing with categorical variables, contingency table is used to determine the relationship between the two variables. Based on the size of the contingency table, the degrees of freedom for the chi-squared distribution are determined as:

degrees of freedom: (rows - 1) * (cols - 1)

The experiment can be described as follows in terms of a p-value and a defined significance level (alpha):

- If p-value <= alpha: significant result, reject null hypothesis (H0), dependent.
- If p-value > alpha: not significant result, fail to reject null hypothesis (H0), independent.

4.1.1 First Hypothesis testing

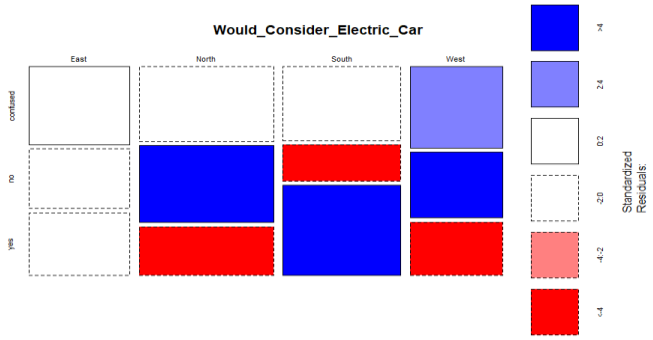
The first hypothesis is tested to determine the existence of a relationship for 3.2.2.1 section.

Null Hypothesis (H0): There is no correlation between the two categorical variables. It means that the choice of considering electric / hybrid car is independent of European regions.
Alternate Hypothesis (H1): There is correlation between the variables. It means that the choice of considering electric / hybrid car is dependent on European regions.

Result
The chi-squared value is 1128.1 with p-value almost equal to 0. Considering the significance level of 5%, the null hypothesis is rejected and it is concluded that European regions have a dependency on their choices of buying electric/hybrid cars in their next purchase.

X-squared	df	p-value
1128.1	6	< 2.2e-16

<Table3: Chi-squared test result for first hypothesis testing>



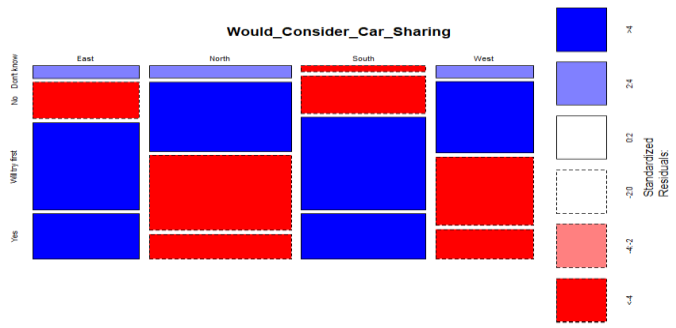
<Figure4: Mosaic plot for showing the standardized residuals of first hypothesis test>

4.1.2 Second Hypothesis testing
The second hypothesis is tested to determine the existence of a relationship for 3.2.2.2 section.
Null Hypothesis (H0): The choice of considering car-sharing option has no correlation with the European regions.
Alternate Hypothesis (H1): European regions have relationship with their choices of considering car-sharing option.

Result
The chi-squared value is 1361.7 with p-value almost equal to 0. Hence, the null hypothesis is rejected and it is deduced that European regions have a dependency on their choices of considering car-sharing option.

X-squared	df	p-value
1361.7	9	< 2.2e-16

<Table4: Chi-squared test result for second hypothesis testing>



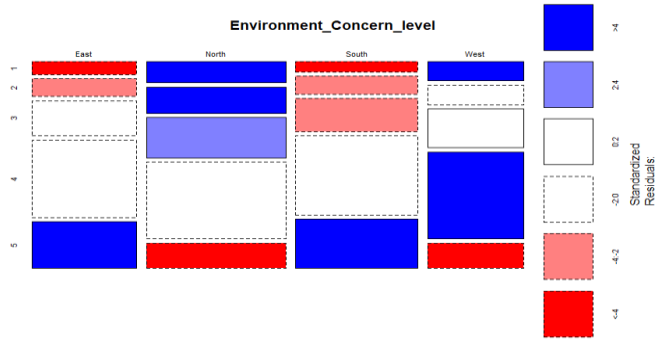
<Figure5: Mosaic plot for showing the standardized residuals of second hypothesis testing>

4.1.3 Third Hypothesis testing:
The third hypothesis is tested to determine the existence of a relationship for 3.2.2.3 section.
Null Hypothesis (H0): The level of concern on environment impacts due to road traffic has no dependency on European regions.
Alternate Hypothesis (H1): The level of concern on environment impacts due to road traffic has dependency on European regions.

Result
The chi-squared value is 816.65 with p-value almost equal to 0. Thus, the null hypothesis is rejected and it is concluded that European regions have a dependency on their grading towards the environmental concern level due to traffic.

X-squared	df	p-value
816.65	12	< 2.2e-16

<Table5: Chi-squared test result for third hypothesis testing>



<Figure6: Mosaic plot for showing the standardized residuals of third hypothesis testing>

V. METHODS USED AND WHY

5.1 Pearson's Chi-squared Test
Chi-squared test is used for testing all the three hypotheses (4.4.1, 4.1.2 and 4.1.3) as all the variables are categorical in nature. This test assumes (the null hypothesis) that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable [2].

- The Chi-Squared test does the test for a contingency table, first calculating the expected frequencies for the groups, then determining whether the division of the groups, called the observed frequencies, matches the expected frequencies. The result of the test is a test statistic that has a chi-squared distribution and can be interpreted to reject or fail to reject the assumption or null hypothesis that the observed and expected frequencies are the same [4] .
- The chi-square test of independence works by comparing the categorically coded data that is collected (known as the observed frequencies) with the frequencies that would expect to get in each cell of a table by chance alone (known as the expected frequencies) [4].

To calculate the chi-squared (c2) statistic the value of

$$\frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

needs to be calculated for each cell in the contingency table and their summation produces the chi-squared value.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The standardized residuals from the test result show the measure of the strength of the difference between the observed and the expected values [5][6].

- If the residual is less than -2, the cell's observed frequency is less than the expected frequency.
- Greater than 2 and the observed frequency is greater than the expected frequency.

Standardized residual = (observed count – expected count) / √expected count [7]

VI. Results and Findings

By opting for an electric car and subscribing to the car-sharing option, European people can reduce the rate of carbon emission to have a lesser effect on the environment. When Europe is divided into four regions (north, south, east and west), it is interesting to see different responses from different regions for the above-mentioned options. The notable findings from the results are:

- Figure 4 of the first hypothesis test depicts that the "yes" cell of the southern region for considering electric car has a residual value of > 4 which means that the observed frequency is higher than the expected frequency. Thereby, the Southern region has shown an inclination for buying electric cars. Moreover, the "no" cell of Northern and Western region has a residual value of <- 4, which means that the observed frequency is less than the expected frequency. Consequently, the Northern and Western region have shown a propensity towards not buying electric cars. Additionally, the Eastern region has shown an average response to all the categories.
- Figure 5 of the second hypothesis test proves that Eastern and Southern regions have shown indication towards considering the car-sharing option and considering to try the option first before subscribing with a similar residual value > 4 for each option. Alternately, both of the regions have shown aversion for the "no" category of car-sharing option with a residual value <-4. Northern and Western regions have shown indication towards not considering the car-sharing option with residual value > 4. These regions have also shown aversion towards considering and trying the option with residual value <-4.
- Figure 6 of the third hypothesis test shows that Eastern and Southern regions have shown a significant level of concern towards environmental impact with a residual value > 4 for level = "5" (Very concerned) and a residual value of <-4 for level = "1" (less concerned). In contrast, Northern and Western regions have shown least concern towards the environment with a residual value of > 4 for level = 1 (less concerned) and a residual value of <-4 for level = 5 (Very concerned).

VII. CONCLUSION

The results of the hypothesis tests conclude that there is a relationship between the European Regions and their propensity to buy an electric car and opting car-sharing method. Consequently, the regions show a propensity towards their level of concern on environmental impact due to traffic and more cars on the road. It is depicted from the results that Eastern and Southern countries show significant concern towards the environment, and it is reflected in their inclination for buying electric cars and car-sharing. On the other hand, Northern and Western countries have shown aversion towards car-sharing and buying electric cars, and by that means shown less concern for the impact of cars on the environment.

References

[1] Christidis, Panayotis. 2016: EU Travel Survey on demand for innovative transport systems. European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/jrc-tem-eu_travel_survey_2014_new_technologies.

[2] Earl T., Petit Y.L. 2019. Transport & Environment (2019) Less (cars) is more: how to go from new to sustainable mobility. [ONLINE] Transport and Environment. Availbale at: https://www.transportenvironment.org/sites/te/files/publications/Less_is_more_4Rs_FINAL%20%281%29.pdf

[3] Archer G., Bondorova B. 2017, Does sharing cars really reduce car use?.[ONLINE]. Transport and Environment. Availbale at: https://www.transportenvironment.org/sites/te/files/publications/Does-sharing-cars-really-reduce-car-use-June%202017.pdf

[4] Brownlee, J. 2019. A Gentle Introduction to the Chi-Squared Test for Machine Learning. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/chi-squared-test-for-machine-learning/.

[5] Kaye, D.H. and Freedman, D.A., 2011. Reference guide on statistics. *Reference manual on scientific evidence*,, pp.211-302.

[6] Urdan, T.C., 2011. *Statistics in plain* English. Routledge.

[7] Stephanie. 2013. Statistics How To. [ONLINE] Available at: https://www.statisticshowto.datasciencecentral.com/what-is-a-standardized-residuals/.