Middle East Technical University

Department of Statistics

2023-2024 Spring

STAT 250 Project Report

14-06-2024

Examining many aspects of transportation in the capital city of Turkiye

ATA ADANUR
BORA ESEN
İREN SU ÇELİK
EMİRHAN KIRAN

Abstract: Ankara, the second most crowded city in Turkey after Istanbul, has various means of transportation that citizen can make use of. Ankara is praised by the citizens until recently because of its accessible and easy transportation and very little traffic, allowing them to travel large distances in short time. However, this perception has changed, and some people believe that it is harder and more inconvenient to use any means of transport in Ankara. This research has been done to find out if transportation has changed or not, if it has changed what are the factors that cause the change. This document analyses transportation in Ankara, how and why it has changed from July 2022 to June 203. In this study, data, that was available for the time interval, are gathered from several resources such as, Ankara Metropolitan Municipality database and EGO database. During this research, a variety of programming languages and visualization tools are made use of like R, Python, Excel and Tableau. This document first provides descriptive statistics and visuals, and then, research questions and analysis followed by results and makes closure with a conclusion.

1.Introduction

The capital city of Turkey, Ankara, has a total population of 5.803.482 (Figure 1) in 2023 and is ranked the second with its population among 81 provinces in Turkey. The population of Ankara increases every year and has now reached its highest with a rise of approximately 1 million in 12 years.

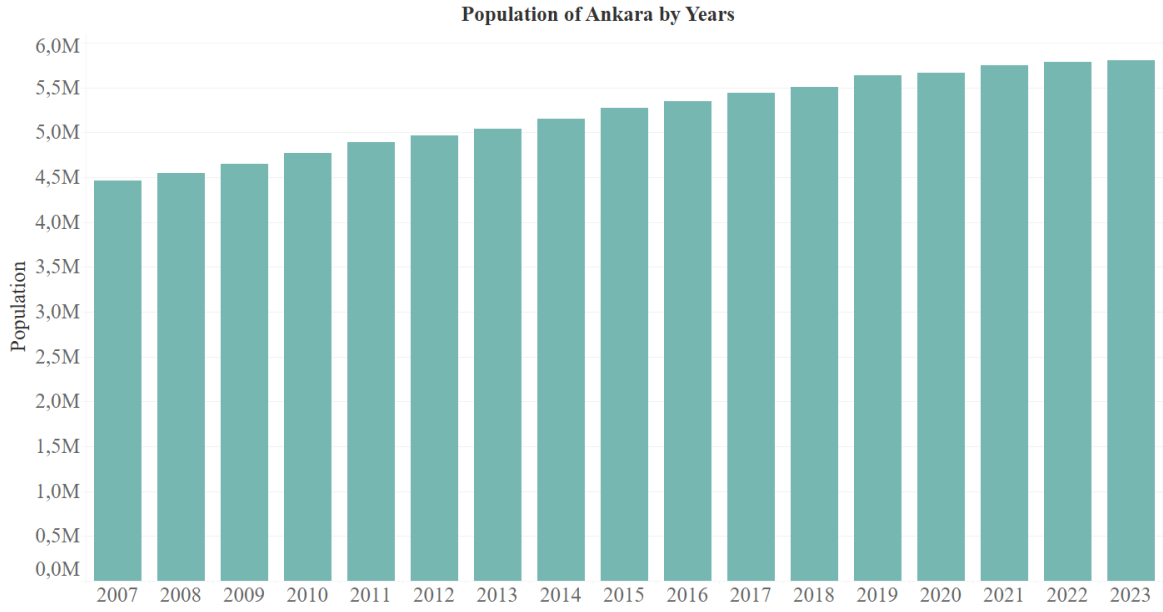


Figure 1. Total population of Ankara from years 2007 to 2023.

The city has a net migration rate of 6.725 per year on average, with a net migration number of 34814 immigrants on average per year. It is believed that the population of Ankara had a rapid climb after the 6th of February earthquake in Hatay because of the immigration it got from Mediterranean and Southeast regions. Even though Ankara has received 232700 immigrants in 2023, which is the highest in 15 years, the net migration is only 23960 people, lower than average net migration, proving the common belief is not right.

Wang et al. (2018) state that the traffic index reflects the state of traffic flow. It is a measurement to evaluate efficiency of the transportation systems, including factors such as vehicle speeds, traffic volumes and accidents (Petrova et al., 2020). The average traffic index of Ankara is 141.93 and the average traffic index of Istanbul is 268.25. For the year 2022, traffic index of Ankara is 148.2 and traffic index of Istanbul is 245.7. In 2023, it is 148.5 for Ankara and 241.8 for Istanbul.

In Ankara, total number of cars per thousand people is 322 in 2023 making it the highest number among other years (Figure 2). Between the years 2022 and 2023 it has increased by 21 cars per thousand people. Total number of vehicles including cars, motorcycles, buses, minibuses, trucks and vans is 2339242 in 2022 and 2479601 in 2023. Every year the number of cars that are newly registered are 87943 on average where total number of cars registered is 94066 in 2022 and 160994 in 2023, the highest in 20 years.

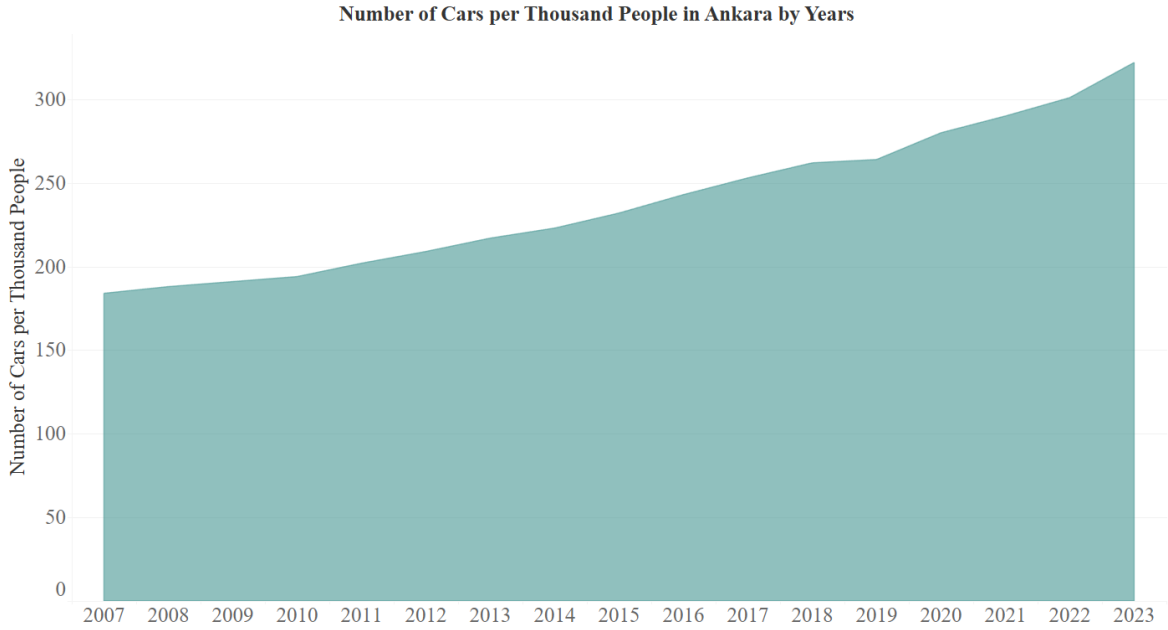


Figure 2. Number of cars per thousand people in Ankara by years from 2007 to 2023.

Personal vehicles are used as a means of transportation in Ankara; moreover, public transportation is also a widely used means of transportation. In the year 2022, number of public transportation passengers is 1393983 on daily average, with a proportion of 48.35% EGO passengers, with 22.17% metro passengers and with 7% Ankaray passengers. Usage of public transportation has increased in 2023 becoming 1476044 passengers on daily average, with a proportion of 49.24% EGO passengers, with 23.53% metro passengers and with 6.5% Ankaray passengers.

Passengers can be identified by the card they use. Senior passengers use +65 card and student passengers use student card. Average number of senior passengers per month is 2883304 in 2022 and 3695783 in 2023. Monthly average number of student passengers is 15159153 in 2022 and 17506481 in 2023. Both number of senior and number of student passengers have increased from 2022 to 2023.

2. Materials and Methods

2.1 Materials

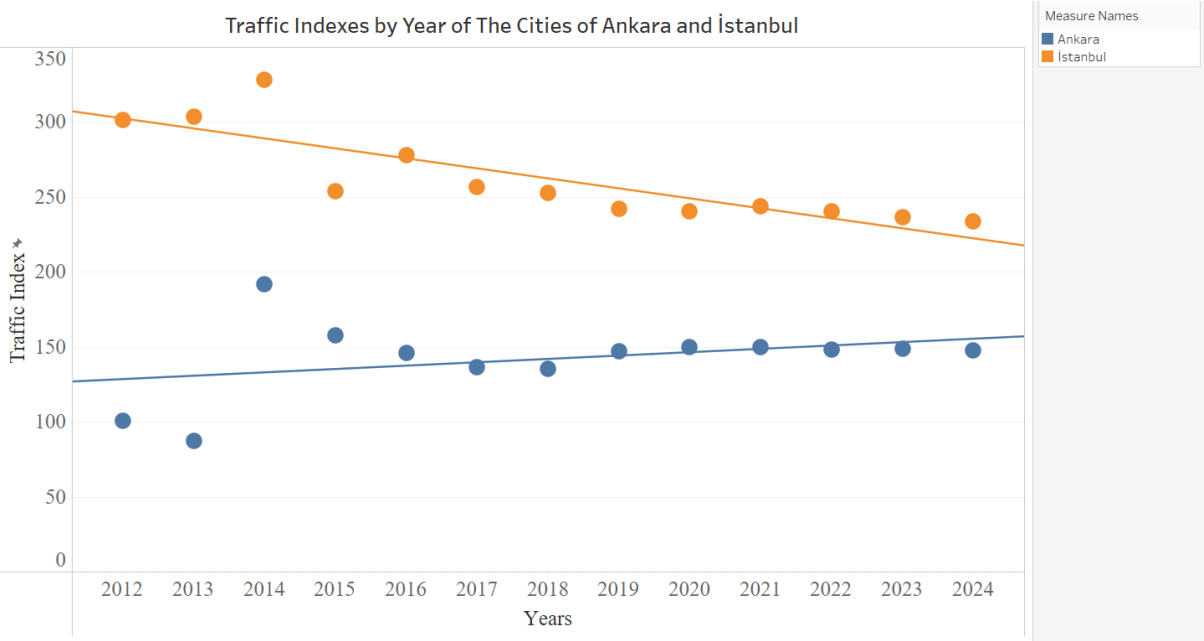
We have done extensive research to find datasets to use in our hypotheses tests and to find answers to our research questions. Population and migration data are derived from TÜİK's Population Statistics Portal. Total number of cars, number of cars registered and total number of vehicles by the type of vehicle data are derived from TÜİK. Public transportation usage according to type of transportation data are derived from EGO. Public transportation usage according to type of card data is derived from Seffaf Ankara. Traffic indexes of Ankara and Istanbul data are derived from Numbeo. We created new data sets in Excel according to the time interval with the columns of the data that we will use. We have excluded one outlier from public transportation usage data and filled empty rows in public transportation usage by type of card data with the means.

2.2 Methodology and Results

Hypothesis Tests in this section and their assumptions are conducted via R. Their visualizations are done in Tableau and tables are created in Word.

a) Two Sample Mean Hypothesis Test Between the Traffic Indexes of Ankara and Istanbul

Dataset Information



For İstanbul, mean is 268.246 with standard deviation of 31.23. Mode is 245.7 and median is 258.7. IQR is 50.45.

For Ankara, mean is 141.93 with standard deviation of 25.44. Median is 147.4. IQR is 13.9.

The first 5 observations of the dataset are given below.

DATE	TRAFFIC INDEX OF ANKARA	TRAFFIC INDEX OF İSTANBUL
2012	100,4	308,1
2013	87,3	310,5
2014	191,6	335,3
2015	157,5	259,3
2016	146,2	284,2

To compare the traffic indexes of Ankara and Istanbul, Two Sample Mean Hypothesis Test seemed appropriate, then we checked the assumptions of the test. X_1, Ankara Traffic Index, has a mean of 141.93 and standard deviation of 25.44. X_2, Istanbul Traffic Index, has a mean of 268.246 and standard deviation of 31.23. Since the populations are independent, independency of samples assumptions is met.

X_1: Traffic Index of Ankara
X_2: Traffic Index of Istanbul

Our null and alternative hypotheses are:

H_0: $\mu_1=\mu_2$

H_1: $\mu_1<\mu_2$

We also need to check variances,

Thus, we'll use F-test to test for differences in variances.

$H0$: $\sigma_1^2=\sigma_2^2$

$H1$: $\sigma_1^2\neq\sigma_2^2$

Since the p-value (0.4887) is greater than 0.05 which is our alpha value, we cannot reject the null hypothesis. Therefore, we can say that the variances of two population are equal. So, we will continue our test with the assumption of equal variances.

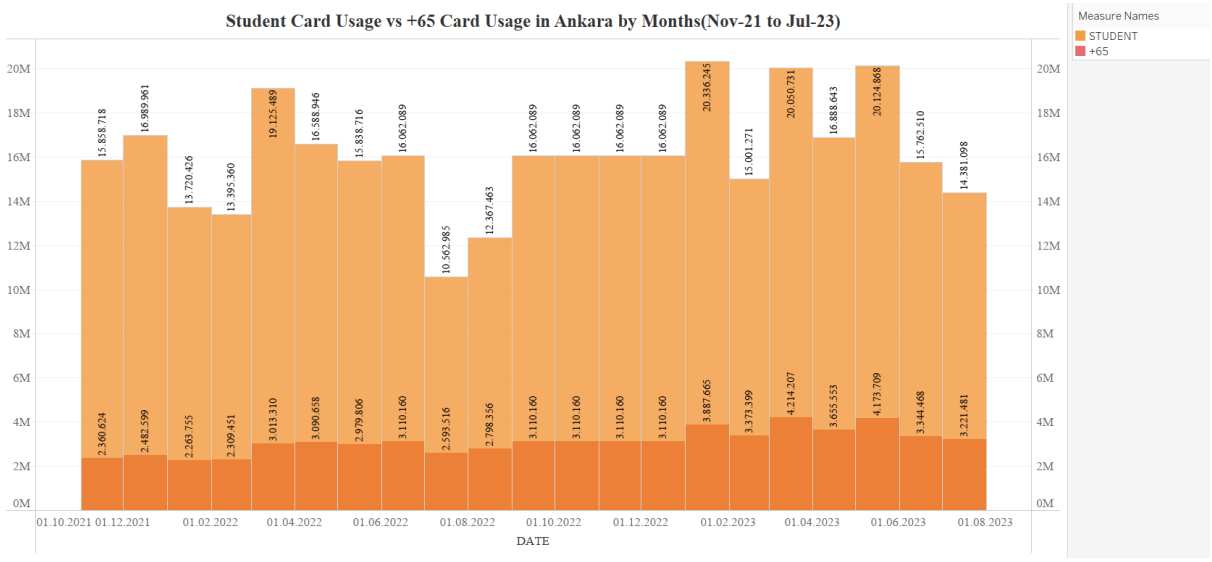
Since the p-value (2.128e-11) is lower than 0.05, we reject the null hypothesis. Therefore, we can say that the mean traffic index of Ankara is not equal to the mean traffic index of Istanbul.

Also, the confidence interval does not include 0. That's why we can reject the null hypothesis.

Therefore, there is enough evidence to say that the mean traffic index of Ankara is smaller than the mean traffic index of Istanbul.

b) Two Proportion Hypothesis Test Between the Proportions of Student Passenger and Senior Passenger

<u>Dataset Information</u>



For Student Card Mean is 16062089.3 with standard deviation of 2474947.34. Mode is 16062089 and Median is also 16062089. IQR is 2248117.5

For +65 Card Usage mean is 3110159.86. Median is 3110160. Mode is also 3110160. Standard deviation is 551936.577. IQR is 662997.5

The first 5 observations of the dataset are given below.

DATE	+65 CARD	STUDENT CARD
Nov.21	2360624	15858718
Dec.21	2482599	16989961
Jan.22	2263755	13720426
Feb.22	2309451	13395360
Mar.22	3013310	19125489

To compare the proportions of the student and senior passengers we used Two Proportion Hypothesis Test. The assumptions of the test ($n1\hat{p}$, $n1(1-\hat{p})$, $n2\hat{p}$, and $n2(1-\hat{p})$ are greater than 5) is met.

X_1: Number of Student Passengers
X_2: Number of Senior Passengers

p_1: the proportion of student passenger
p_2: the proportion of senior passenger

Our null and alternative hypotheses are:

H_0: $p_1=p_2$

H_1: $p_1> p_2$

The p-value of the test is less than 2.2e-16, which is lower than the significance level (α) of 0.05. Consequently, we reject the null hypothesis at the 5% significance level. This result implies that there is enough evidence to conclude that the population proportion of student passengers is greater than the population proportion of senior passengers.

c) Simple Linear Regression of Number of Vehicles Registered and Total Population

Dataset Information

For Total Population, mean is 5217141 with standard deviation of 447754.588. Median 5270575 IQR is 819894.5,

For Number of Vehicles Registered, the mean is 87943.05 with standard deviation of 21733.81 Median is 85892. 2023 is outlier with 160994. IQR is 21875.5

The first 5 observations of the dataset are given below.

DATE	# OF VEHICLES	TOTAL POPULATION
2012	85897	4965542
2013	88632	5045083
2014	75263	5150072
2015	99243	5270575
2016	99898	5346518

To understand the relationship between the dependent variable, number of cars registered, and the response, total population, linear regression model is appropriate. Since only one independent variable is used, we created simple linear regression model.

At first, an extreme value is spotted which lies on Cook's distance in Residuals vs Leverage Plot. To achieve a better fit of the model, the extreme value is removed, and assumption tests are done again.

The scatterplot of the independent variable, population, and response, number of vehicles registered, shows a linear relationship.

In Q-Q Plot, residuals follow the straight dashed line which indicates that residuals are normally distributed.

In Durbin-Watson Test, the null hypothesis states that the errors are independent. Since the p-value equals to 0.056 which is greater than 0.05, we fail to reject the null hypothesis and we can say that the errors are independent.

In Scale-Location Plot, there is a horizontal line with equally spread points which is a good indication of homoscedasticity, homogeneity of variance of the residuals.

Since all the assumptions of simple linear regression are met, we can conduct the simple linear regression model.

Our null and alternative hypotheses are:

H_0 : β_0 is not significant (β_0=0)

H_1 : β_0 is significant ($\beta_0 \neq 0$)

The p-value for the significance of the coefficients β_0 is 0.3119 which is greater than 0.05 so we fail to reject the null hypothesis and we can say that β_0 is insignificant.

H_0 : Model is not significant (β_1=0)

H_1 : Model is significant ($\beta_1 \neq 0$)

The p-value for the significance of the coefficients $\hat{\beta}_1$ is 0.0228 it is less than 0.05 so we reject the null hypothesis and we can say that $\hat{\beta}_1$ is significant. Therefore, we can say that the model is significant.

The confidence interval for $\beta 1$ is (1.6582, 19.0955) which does not contains 0 so the $\beta 1$ and the model is significant.

When we look at the output, R^2 is 0.3002. So, the regression model shows that approximately 30% of the total variation in number of cars registered is explained by total population.

Correlation coefficient is 0.5479 so we can say that there is a medium positive relationship between population and number of vehicles registered.

The least squares estimation equation is (Figure 3),

$\hat{y}$=(-6.249e+04)+(2.893e-02)*x

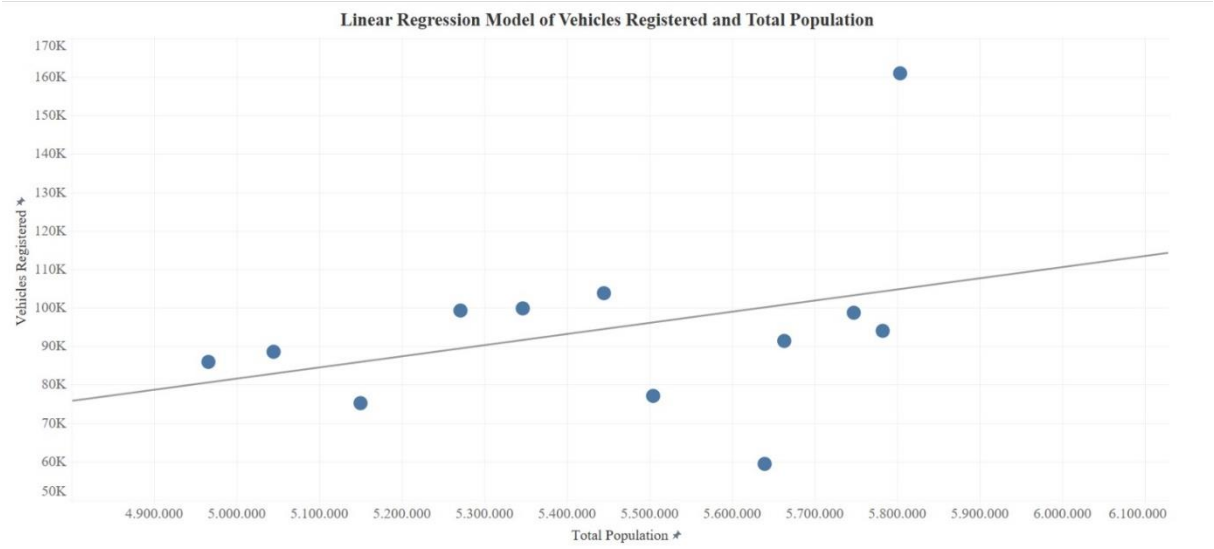


Figure 3. Linear Regression Model of Vehicles Registered and Total Population.

d) Multiple Linear Regression of Total Public Transportation Usage, Standard Card Usage and Total Number of Vehicles

Dataset Information

The first 5 observations of the dataset are given below.

DATE	Total Number of Vehicles	Total Public Transportation Usage	Standard Card Usage
Jul.22	2339242	31034666	13351049
Aug.22	2349305	35500416	15187767
Sep.22	2357360	38878392	14484273
Oct.22	2366078	38878392	14484273
Nov.22	2376573	38878392	14484273

For Total Public Transportation Usage mean is 38878392.14. Median is 38878392. Mode is 38878392. Standard deviation is 4079318. IQR is 24722796.5

For Total Number of Vehicles mean is 2401397.83 with standard deviation of 46617.2524. Median is 2395467 and IQR is 77464.5

For Standard Card Usage mean is 14726143.3 with standard deviation of 1038662.15. Median is 14726143.3 and IQR is 1416618.

To understand the relationship between the dependent variable, total public transportation usage, and the independent variables, standard transportation card usage (x_1) and total number of vehicles (x_2) linear regression model is appropriate. Since there are more than one independent variable, we created multiple linear regression model.

At first, we have created a multiple linear regression model with 3 independent variables, which are standard card usage, total vehicles and fuel prices and checked the assumptions. However, summary of the model showed that fuel prices is an insignficant variable. Since we have already checked the assumptions, we removed the insignificant variable to obtain a new model.

Shapiro-Wilk Test with a p-value that is equal to 0.3227 indicates that total public transportation usage is normally distributed.

Residuals vs Fitted Plot shows a horizontal line without any distinct patterns which indicates that there is a linear relationship between the predictor, and the outcome variables.

In Q-Q Plot, residuals follow a straight line on the dashed line indicating that residuals are normally distributed.

In Durbin-Watson Test, the null hypothesis states that the errors are independent, since the p-value equals to 0.704 which is greater than 0.05, we fail to reject the null hypothesis and we can say that the errors are independent.

In Scale-Location Plot, the line is roughly horizontal, and the residuals are mostly, randomly scattered which indicates that homogeneity of variances of the residuals assumption is met.

According to the VIF (Variance Inflation Factor) results, since the value is 1.1421, which is smaller than 5, there is no multicollinearity. So, there is no near-linear dependence among the regression variables, net migration and number of vehicles registered.

Since this fit meets all the assumptions, no transformation is needed to meet assumptions. However, we have done transformation on an independent variable to obtain a better model fit. When sqrt transformation is applied on total vehicles, Adjusted R^2 of the model has increased from 0.85 to 0.861.

Our null and alternative hypotheses are:

H_0 : β_0 is not significant (β_0=0)

H_1 : β_0 is significant ($\beta_0 \neq 0$)

The p-value for the significance of the coefficients β_0 is 0.00734 which is lower than 0.05 so we reject the null hypothesis and we have enough evidence to say that β_0 is significant.

H_0 : Model is not significant ($\beta_1 = \beta_2 = 0$)

H_1 : Model is significant (At least one of them is not equal to 0)

<u>For coefficient β_1:</u>

The p-value for the significance of the coefficients β_1 is 0.00144 which is smaller than 0.05 so we reject the null hypothesis and there is enough evidence to say that β_1 is significant.

<u>For coefficient β_2:</u>

The p-value for the significance of the coefficients β_2 is 0.0094 which is smaller than 0.05 so we reject the null hypothesis and there is enough evidence to say that β_2 is significant.

The p-value of the model is 0.000967 which is lower than 0.05, also the coefficients β_1 and β_2 are significant. So, we can say that the model is significant.

When we look at the output, R^2 is 0.886. This means that the regression model shows that approximately 88.6% of the total variation in total public transportation usage of Ankara is explained by standard card usage and total number of vehicles.

The least squares estimation equation is,

$\hat{y}$ = (-2.771e+08) + (3.391e+00) * x1 + (1.724e+05) * x2

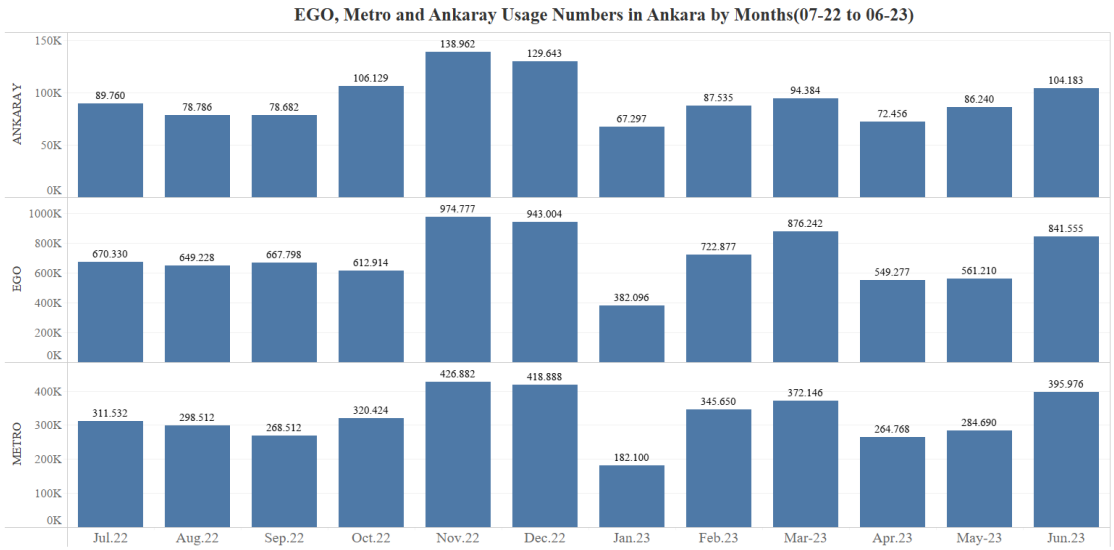
Mean of response equals to -2.771e+08 when x_1= x_2=0. Since 0 is not in the range of x_2, this does not have any practical interpretation.

Expected change in total public transportation usage (y) per unit change in standard card usage (x1) when x_2 is held constant equals to 3.391e+00.

Expected change in total public transportation usage (y) per unit change in total number of vehicles (x2) when x_1 is held constant equals to 1.724e+05.

e) One-Way Anova

Dataset Information



For EGO mean is 704275,667 with std. Dev. Of 176088,426. Median is 669064. IQR is 271836,5

For Metro mean is 324173.333 with std. Dev of 71758.1122. Median is 315978. IQR is 107460

For Ankaray mean is 94504.75 with std. Dev. of 21956.9697. Median is 88647.5. IQR is 26422

The first 5 observations of the dataset are given below.

DATE	EGO	METRO	ANKARAY	OTHER	TOTAL
Jul.22	670330	311532	89760	388133	1369995
Aug.22	649228	298512	78786	375709	1323449
Sep.22	667798	268512	78682	393903	1330213
Oct.22	612914	320424	106129	407431	1340769
Nov.22	974777	426882	138962	536591	1938250

To compare the means of number of EGO, Metro and Ankaray passengers, One-way Anova Test is appropriate, thus we checked the assumptions of the test. Individual Shapiro-Wilk Normality Tests were conducted to check the normality assumption.

p-value of Shapiro-Wilk Test for EGO is 0.7968, for Ankaray is 0.226, for Metro is 0.8393. Since all the p-values are greater than 0.05, we can assume that populations are normally distributed.

The samples are independent from one another.

We conducted pairwise F-tests to check if the variances are equal. From the tests conducted, we couldn't reject the null hypotheses, so, the assumption of equal variances ($\sigma_1^2=\sigma_2^2=\sigma_3^2$) is met.

Our null and alternative hypotheses are:

$H_0: \mu_1=\mu_2=\mu_3$

H_1: At least one mean is different.

The very low p-value (2.67e-14), which is much less than the typical significance level of 0.05, which indicates that the differences between the group means are statistically significant.

The F value of 93.19 indicates that the between-group variability is much larger than the within-group variability, suggesting that the group means are significantly different from each other.

Since the ANOVA test indicates significant differences between the group means, we need to perform a post-hoc test (such as Tukey's HSD) to determine which specific groups' means are significantly different from each other.

According to Tukey's HSD, all p-values are lower than 0.05, indicating that all means are different from each other. To be precise; the pairwise mean difference between EGO and Ankaray types is 609770,9. The pairwise mean difference between Metro and Ankaray types is 229668,6. The pairwise mean difference between Ego and Metro types is 380102,3.

3. Conclusion

In this document, transportation in Ankara is studied. From the study, the truth that Istanbul has worse traffic than Ankara is determined. Even tough Ankara does not have as much traffic as Istanbul, traffic in the capital city is increasing year by year. On the contrary to the common belief that senior citizens use public transportation more than anyone, this study has proven that students use public transportation much more than senior citizens. Moreover, the study shows that some means of public transportation is used more than others in Ankara. EGO is the most used mean of public transportation in Ankara among Metro and Ankaray, Metro is the second most used and Ankaray is the least used among three. Additionally, the models obtained from the study can be useful to make regulations about future development plans. The results of the single regression model can be convenient for making estimations about the number of vehicle registrations. For Ankara Metropolitan Municipality, results of the multiple regression model can be a reliable model to predict the possible future of public transportation usage and to improve public transportation systems in Ankara.

References

Ankara Metropolitan Municipality Şeffaf Ankara Open Data Platform. (2022). *Transportation* [Data file]. Retrieved from https://seffaf.ankara.bel.tr/ (in Turkish).

EGO Head Office. (2020). *Number of Daily Service and Passengers* [Data file]. Retrieved from https://www.ego.gov.tr/tr/sayfa/2280/gunluk-yolcu-sayilari2022. Accessed: 15/04/24. (in Turkish).

EGO Head Office. (2023). *09 March 2020 - 21 - 27 August 2023 Number of Public Transportation Passengers (Table-1)* [Data Set]. https://www.ego.gov.tr/dosya/indir/30865.pdf (in Turkish).

Numbeo. (2012). *Asia: Traffic Index by City 2012* [Data file]. Retrieved from https://www.numbeo.com/traffic/region_rankings.jsp?title=2012-Q1&displayColumn=0&region=142. Accessed: 16/04/24.

Petrova, T., Grunin, A., & Shakhbazyan, A. (2020). Integral Index of Traffic Planning: Case-Study of Moscow City's Transportation System. *Sustainability*, *12*(18), 7395. https://doi.org/10.3390/su12187395

Turkish Statistical Institute Central Distribution System. (2016). *Number of Motor Land Vehicle* [Data file]. Retrieved from https://biruni.tuik.gov.tr/medas/?kn=89&locale=tr (in Turkish).

Turkish Statistical Institute Central Distribution System. (2016). *Number of Motor Land Vehicles Registered* [Data file]. Retrieved from https://biruni.tuik.gov.tr/medas/?kn=89&locale=tr (in Turkish).

Turkish Statistical Institute Central Distribution System. (2016). *Number of Cars per Thousand People* [Data file]. Retrieved from https://biruni.tuik.gov.tr/medas/?kn=89&locale=tr (in Turkish).

Turkish Statistical Institute Data Portal. (2023). *Databases: Motor Land Vehicle Statistics* [Data file]. Retrieved from https://data.tuik.gov.tr/Kategori/GetKategori?p=Ulastirma-ve-Haberlesme-112. (in Turkish).

Turkish Statistical Institute Population Statistics Portal. (2023). *Population by Gender* [Data file]. Retrieved from https://nip.tuik.gov.tr/. Accessed: 21/04/24. (in Turkish).

Turkish Statistical Institute Population Statistics Portal. (2023). *Interprovincial Migration* [Data file]. Retrieved from https://nip.tuik.gov.tr/?value=IllerArasiGoc. Accessed: 21/04/24. (in Turkish).

Wang, W., Guo, R., & Yu, J. (2018). Research on road traffic congestion index based on comprehensive parameters: Taking Dalian city as an example. *Advances in Mechanical Engineering/Advances in Mechanical Engineering*, *10*(6), 168781401878148. https://doi.org/10.1177/1687814018781482