MIDDLE EAST TECHNICAL UNIVERSITY

FACULTY OF ARTS AND SCIENCES

STATISTICS

FACTORS AFFECTING HAPPINESS

Group 9

ZEYNEP FENERCİOĞLU 2561264

BAŞAK UĞURLU 2561546

DAMLA BAŞARMIŞ 2510048

Introduction

In this project, “Factors Affecting Happiness” have been analyzed. Observed variables from the dataset are the Healthcare Index, Air Quality Index, Green Space Area, Cost of Living, and Traffic Density. Research questions that were created to analyze the relationship between the Happiness Score and variables are:

1. Is the average happiness score significantly different from the hypothesized value 5?
2. Is there a significant difference between the means of happiness scores of cities with high traffic density and medium traffic density?
3. Do countries with air quality above %60 degrees have higher happiness rates?
4. Is there a significant difference in the proportion of cities with a high healthcare index between cities with high green space area and low green space area?
5. How does Healthcare Index affect Happiness Score?
6. What Air Quality Index and Healthcare Index affect Happiness Score?
7. Do means of happiness scores differ depending on the healthcare index?

Statistical methods that were used in this project are inferences about the mean, comparisons of means, inferences about proportion, comparisons of proportion, one-way or two-way ANOVA and multiple comparisons, and simple and multiple linear regression. After exploratory analysis, we have decided on the correct statistical tests, checked the requirements and met the assumptions. Finally we made inferences to our research questions from the test results.

Data Source: https://data.world/bareche/us-happiness-data/workspace/file?filename=happiness+data.csv

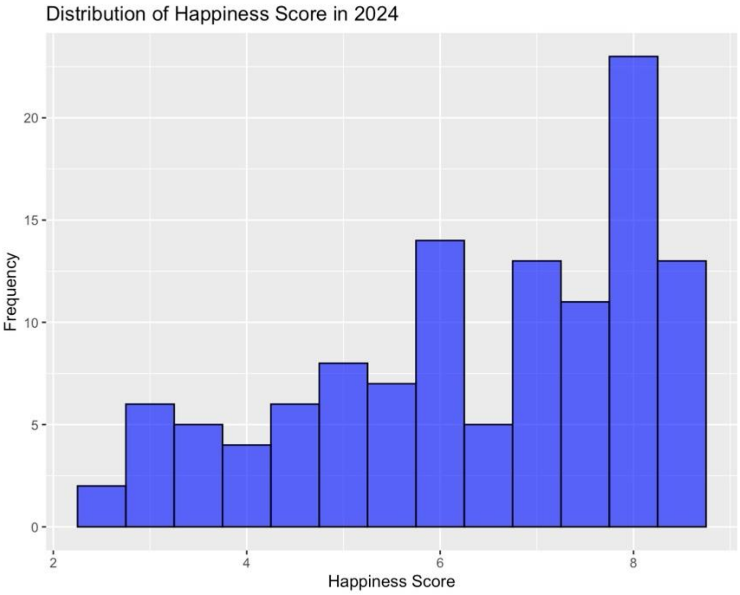
Analysis

1) **Inferences About Mean /One Sample Hypothesis Testing:**

Is the average happiness score significantly different from the hypothesized value 5?

Performing Exploratory Data Analysis:

The mean happiness score is 6.37265. It is raging from 2.5 to 8.6. It is quantitative.



Assumptions:

First, we need to check normality of happiness score. To check the normality, we use Shapiro test. Since its value is less than 0.05, normality assumption is not met for happiness score. We should use a non-parametric test instead of the one-sample t-test.

Performing Wilcoxon Signed-Rank Test:

- Null hypothesis: The median happiness score is equal to 5.
- Alternative hypothesis: The median happiness score is not equal to 5.

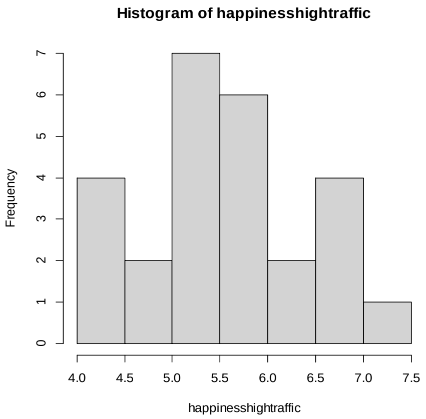
The calculated p-value is 1.068e-11. Thus, we reject the null hypothesis, and we can say that the median happiness score is significantly different from 5.

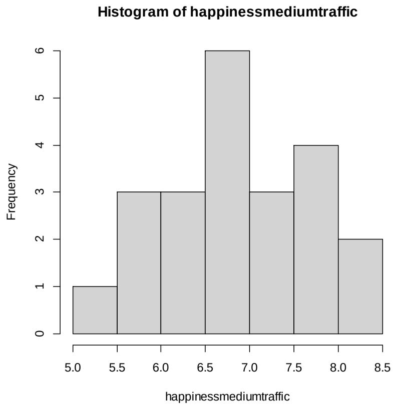
2) **Comparisons of Means/Two Sample Hypothesis Testing :**

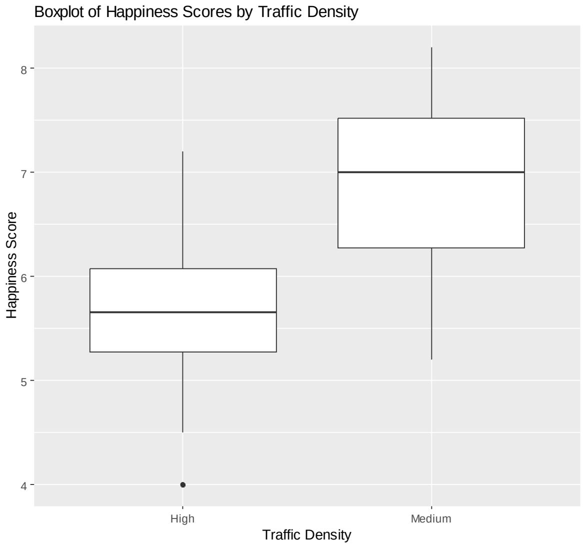
Is there a significant difference between the means of happiness scores of cities with high traffic density and medium traffic density?

Performing Exploratory Data Analysis:

There are 26 observations for happiness scores of high-traffic density cities. The median is 5.65. There are 22 observations for happiness scores of medium-traffic density cities. The median is 7. Distributions are shown below.







Assumptions:

Samples are independent there is no connection between cities with medium traffic density and high traffic density.We don’t know the population standard deviations, we do not know if the happiness score population is normally distributed, and the sample sizes are smaller than 30. Therefore, we use the Mann-Whitney- Wilcoxon test since the requirement of normal distribution for the t-test is not met. Mean ranks approximate the median, so we are testing the median differences. We check whether the distributions are identical so that we can compare medians. As seen from the histograms, they have similar shapes so it is appropriate to use this test.

Performing the Mann-Whitney-Wilcoxon Test:

- Null hypothesis: The two groups (happiness scores of cities with medium and high traffic density) are sampled from populations with identical distributions.
- Alternative hypothesis: The two groups are sampled from populations with different distributions.

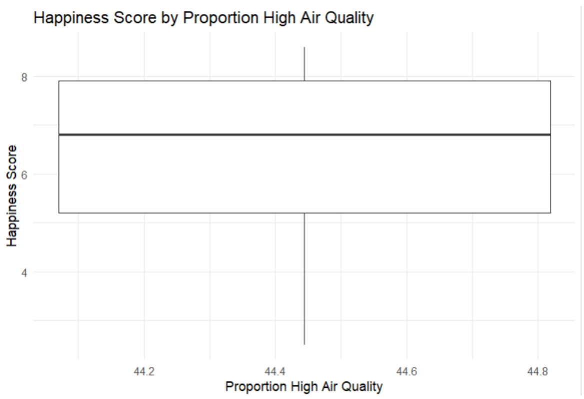
The calculated p-value is 0.0000255. Thus, we reject the null hypothesis, and we can say that happiness scores from cities with high and medium traffic density have different population distributions. Therefore we can say that their medians are significantly different.

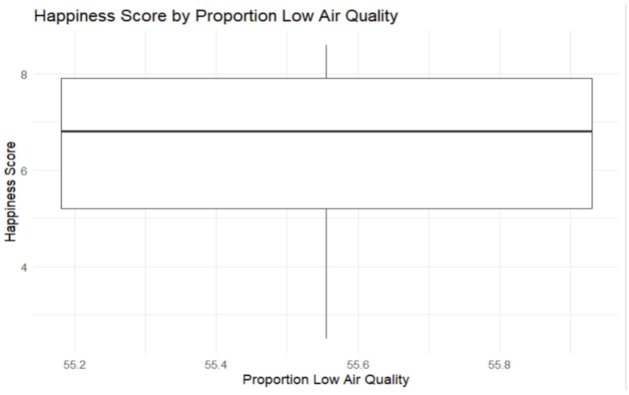
3) Inferences About Proportion/One Sample Hypothesis Test:

Do countries with air quality above %60 degrees have higher happiness rates?

Performing Exploratory Data Analysis:

The proportion of cities with an air quality index below %60 is %55.55 in the dataset. The proportion of cities with an air quality index above %60 is %44.44 in the dataset. The average happiness score of cities with an air quality index below %60 is 7.4969. The average happiness score of cities with an air quality index above %60 is 4.967. There are 52 observations that the Air Quality Index is more than or equal to 60. There are 65 observations that the Air Quality Index is less than 60.





Assumptions:

Since each observation or sample is independent of the other, there is no relationship between cities with low air quality and cities with high air quality. Additionally, the data comes from a random sample. Since np>=10 and n(1-p)>=10, a normal approximation method can be used. But; We don't know how the happiness score is normally distributed. Also, Central Limit theorem is used if the sample size is more than 30 and our sample size is more than 30. Therefore, central limit theorem can be used. The Shapiro-Wilk Test is suitable for our data because it is quite sensitive for small and medium-sized samples.Since standard deviations of the population is not known, the t-test can be used.

Performing Shapiro Wilk Test:

- Null Hypothesis: The sample data follow a normal distribution
- Alternative Hypothesis: The sample data does not follow a normal distribution.

The result of the Shapiro-Wilk Test indicates that W is approximately 0.84205. Also, the p-value is associated with 0.0000008124. Because of the p-value is less than 0.05, we reject the null hypothesis. We conclude that data is not normally distributed so it can be approached to normal distribution.

Performing t-test:

- Null Hypothesis: Countries with air quality above %60 degrees have higher happiness rates
- Alternative Hypothesis: Countries with air quality less %60 degrees have higher happiness rates

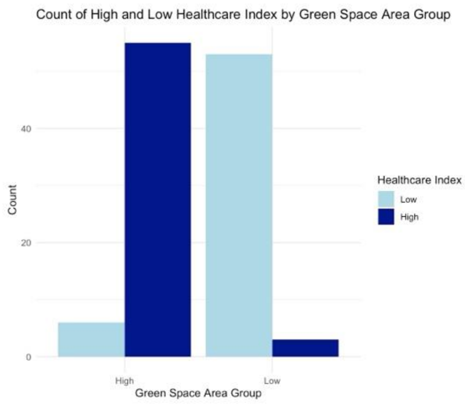
The result of the t test, p value is 1 so it is greater than significance level. We can say that fail to reject null hypothesis.

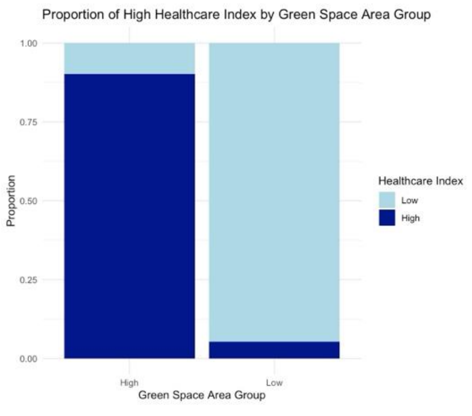
4) Comparisons Of Proportions /Two-Sample Hypothesis Testing:

Is there a significant difference in the proportion of cities with a high healthcare index between cities with high green space areas and low green space areas?

Performing Exploratory Data Analysis:

The mean of the healthcare index is 80.59829. We will split the data based on the median green space area. The median green space area is 55. Low green space area is below 55 and high green area is above 55. The visualizations indicate that cities with high green space areas tend to have a significantly higher proportion and count of cities with a high healthcare index compared to cities with low green space areas.





Assumptions:

First we calculate the sample sizes and the number of successes and failures in each group. We need to ensure that the expected frequencies of successes and failures in each group are at least 5. For high green space area successes are 55 and failures are 6. Low green space area successes are 3 failures are 53. The assumption of having at least 5 expected frequencies of successes and failures in each group is not met. We should use an alternative test that does not rely on this assumption. The

Fisher's Exact Test is a suitable non-parametric alternative for comparing proportions when the sample size is small and the expected frequencies are less than 5.

Performing Fisher's Exact Test:

•Null hypothesis: The proportion of cities with a high healthcare index is the same in both groups.

•Alternative hypothesis: The proportion of cities with a high healthcare index is different in both groups.

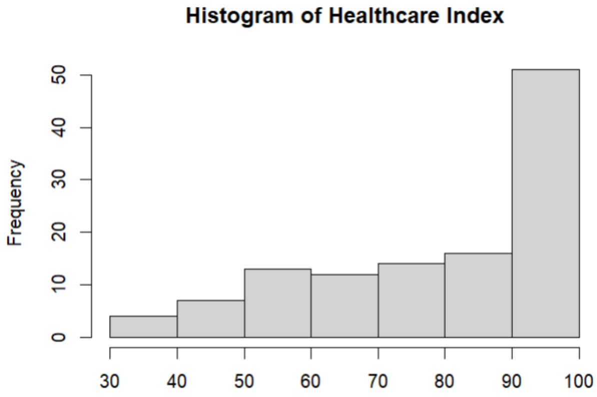
The calculated p-value is 2.2e-16, which is smaller than 0.05. Thus, we reject the null hypothesis. There is a significant difference in the proportion of cities with high healthcare index between high and low green space area groups.

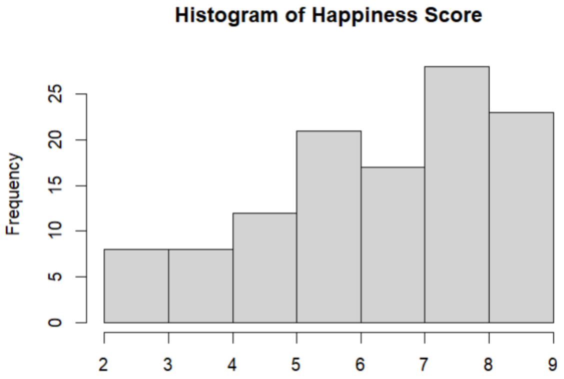
5) Simple Linear Regression and Multiple Linear Regression:

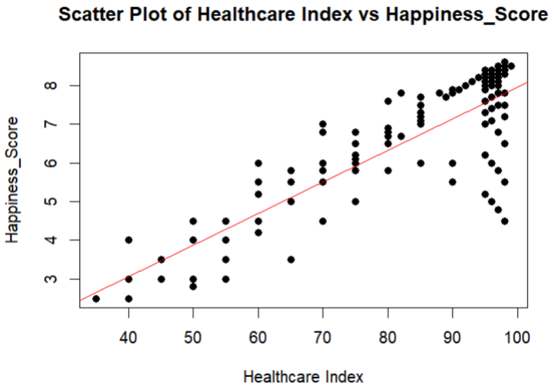
How does the Healthcare Index affect the Happiness Score?

Performing Exploratory Data Analysis:

There are 117 variables for the Healthcare index. There are 117 variables for the Happiness Score. Histogram of Healthcare Index and Happiness score have left skewed distribution. Therefore, density is higher on the left side.







Assumptions:

-Linearity:

The relationship between the predictors and the response variable should be linear.We should create a scatter plot and add a regression line to visualize the linear relationship. If it is close to 1,-1, we can see that it is a strong relationship. If it is close to 0, there is a weak relationship. Our correlation coefficient is 0.8470444, so there is a strong positive relationship.

-Independence:

Tests such as the Durbin-Watson test can be used to evaluate this assumption.

Durbin-Watson Test

The statistics of Durbin Watson test is 0.6766644 and p value is 0.533.Because of the (p value>0.5), It is assumed that there is no autocorrelation between the terms. As a result, it can be said that there is no significant autocorrelation between the error terms in simple linear regression model and the independence assumption is met.

-Homoscedasticity:

The graph of homoscedasticity with a smoothly running graph may indicate that our model satisfies the homoscedasticity assumption. Representing homoscedasticity with a smoothly running graph may indicate that our model satisfies the homoscedasticity assumption. This situation shows that the error terms have a constant variance at the levels of the independent variables and the distribution of the errors does not change

-Normality of Residuals:

Q-Q Plot

A straight line forms in the Q-Q graph, which indicates that the error periods comply with the normal distribution. A solid line indicates that the observed error terms fit well within a normal distribution. Thus, our regression model has a normal distribution.

```
Call:
lm(formula = Happiness_Score ~ Healthcare_Index, data = df)

Coefficients:
     (Intercept)  Healthcare_Index  
          97.322            -1.527  
```

β0=97.322

β1=-1.527

Model: Happiness_Score=97.322−1.527·Healthcare_Index

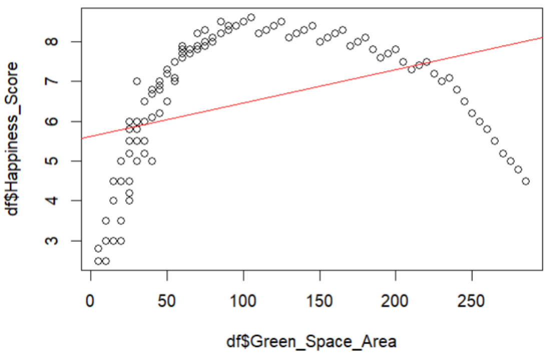
What Air Quality Index and Healthcare Index affect Happiness Score?

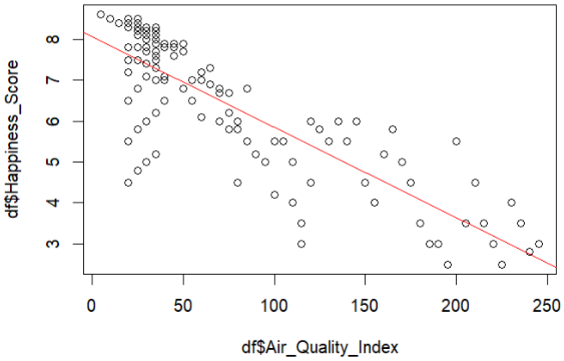
Performing Exploratory Data Analysis:

The mean of the Air Quality Index is 76.58, median is 45, the max value is 245, and min value is 5.The mean of the Green Space Area is 89.87, the median is 55, the max value is 285, and min value is 5.

Assumptions:

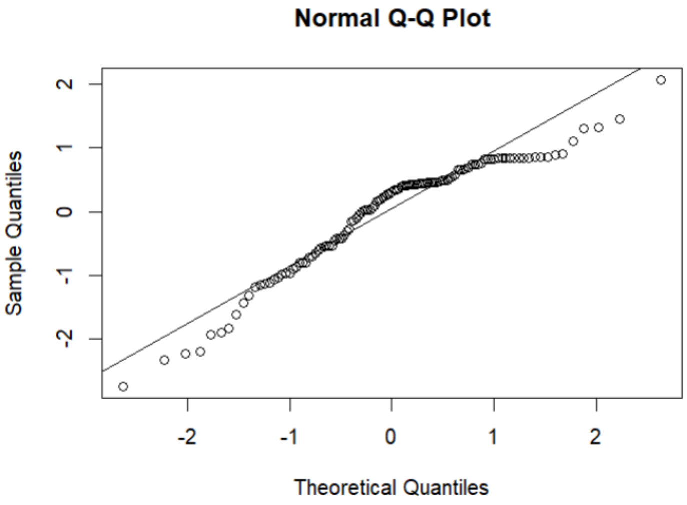
-Linearity:





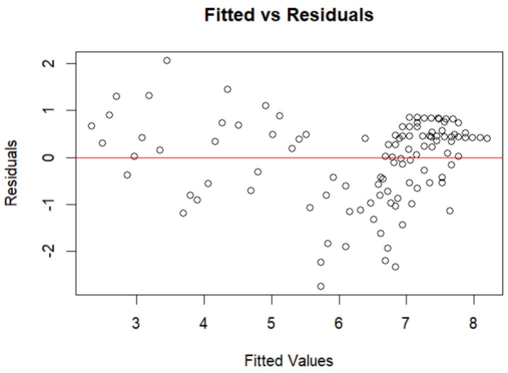
The linearity assumption indicates that the relationship between the independent variables and the dependent variables is linear. There is a positive weak relationship between Green Space Area and Happiness Score because the correlation coefficient is 0.3925. There is a strong negative relationship between the Air Quality Index and Happiness Score because the correlation coefficient is -0.8345

-Normality of Residuals:



The Q-Q plot is often used to visualize how well a normal distribution of error terms fits another distribution. The graph has a straight line and it shows that there is a normal distribution between the two distributions.

- Homoscedasticity:



The distribution graph is straight and a constant line forms. It shows that homoscedasticity is met. That is, predictions of this model have a similar variance for different independent variable values. Also, having homoscedasticity increases the accuracy and robustness of the regression model.

-Multicollinearity

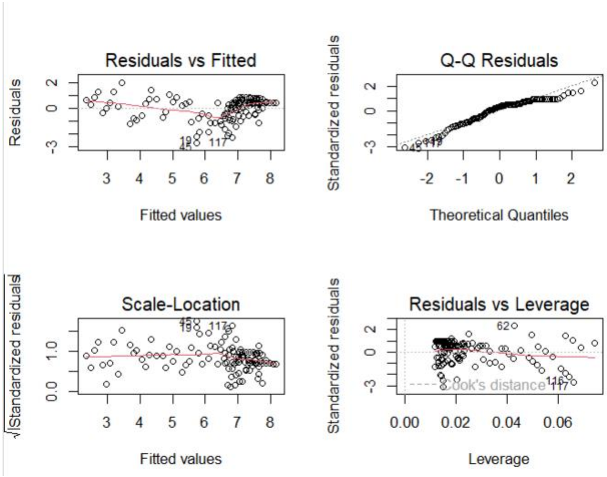
There is a moderate multicollinearity because the variation inflation factor is 1.715366. The VIF values are smaller than 5 so there is no multicollinearity problem.

-Independence of Errors:

Durbin-Watson Test:

The Durbin-Watson test is often used to check for autocorrelation of errors in the regression model.As a result of the Durbin-Watson test, DW = 0.864332. This indicates a positive autocorrelation. The p-value is 8.411e-11 and smaller than the general alpha value 0.05. It indicates that the null hypothesis is rejected.

Also, The alternative hypothesis test is True autocorrelation is greater than 0. Therefore, the result shows that there is positive autocorrelation so This violates the independence of errors assumption. However we should assume errors are independent.



```
Call:
lm(formula = Happiness_Score ~ Air_Quality_Index + Green_Space_Area, 
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7282 -0.5587  0.3051  0.6553  2.0518 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        8.883840   0.241746  36.749  < 2e-16 ***
Air_Quality_Index -0.026509   0.001685 -15.731  < 2e-16 ***
Green_Space_Area  -0.005353   0.001351  -3.963 0.000129 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8934 on 114 degrees of freedom
Multiple R-squared:  0.7332,	Adjusted R-squared:  0.7285 
F-statistic: 156.7 on 2 and 114 DF,  p-value: < 2.2e-16
```

$\beta 0=8.883840$

$\beta 1=-0.026509$

$\beta 2=-0.005353$

All p-values are lower than 0.001, parameters of estimation are significant.

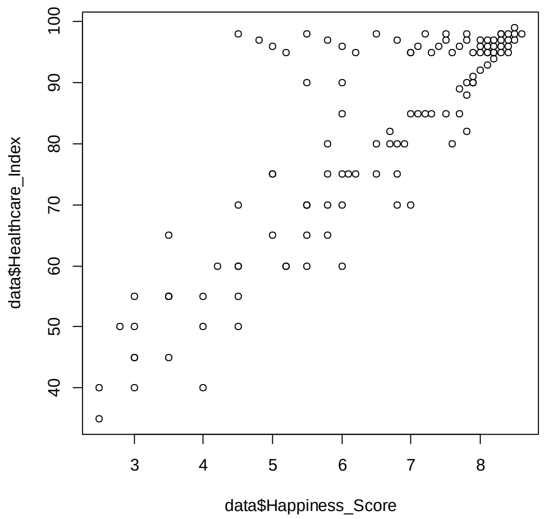
Fitted Model: Happiness Score= 8.883840-0.026509(Air Quality Index)-0.005353(Green Space Area)

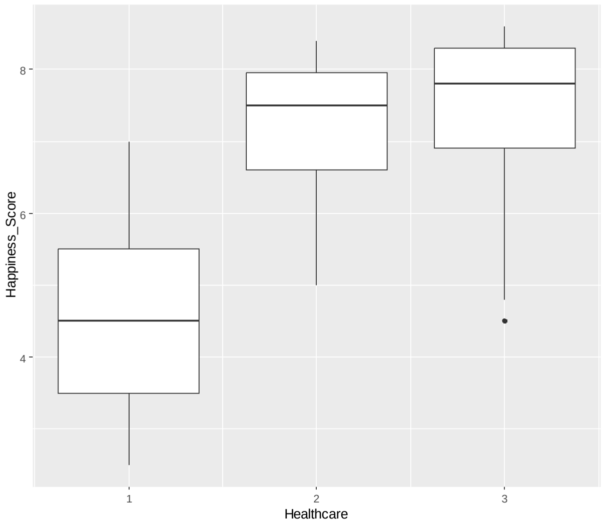
6) One-Way or Two-Way ANOVA and Multiple Comparisons

Do means of happiness scores differ depending on the healthcare index?

Performing Exploratory Data Analysis:

There are 117 observations. The mean happiness score is 6.37265. The mean of the healthcare index is 80.59829. Since both are quantitative, we grouped the healthcare index into 3 groups where each group has the same number of observations, 39. Below 75, 75-95, Above 95. In the scatter plot, we see that points are partially spread. In the box plot, we see that the happiness score of cities with low healthcare indexes has a lower median than that of cities with higher healthcare indexes. Thus, we wanted to check if the differences are significant.





Assumptions:

Populations and cases within each sample are independent. They do not affect each other.We do not know that populations are normally distributed. Each sample size is bigger than 30; we can use the central limit theorem and approximate to normal distribution. By Bartlett's test of homogeneity of variances, variances are equal, with the p-value being 0.1787 we fail to reject the null hypothesis that the samples have equal variances. We can use One-Way ANOVA.

Performing One-Way ANOVA:

•Null hypothesis: Population means of happiness scores from different healthcare index groups are equal.

•Alternative hypothesis: Population means of happiness scores from different healthcare index groups are not equal.

The calculated p-value is 2e-16. We reject null hypothesis and conclude that population means are significantly different.

To see which ones are different, we used Tukey’s procedure of multiple comparison:

For Ho: 2-1= 0 p-value is 1e-04, population means of the 1st and 2nd groups are significantly different. 2nd group has a higher population mean.

For Ho: 3 - 1= 0 p-value is 1e-04, population means of the 1st and 3rd groups are significantly different. 3rd group has a higher population mean.

For Ho: 3 - 2= 0 p-value is 0.759, and the population means of the 2nd and 3rd groups are not significantly different. 2nd and 3rd groups can have equal population means.

Results

1) **Is the average happiness score significantly different from the hypothesized value 5?**

The analysis aimed to determine whether the average happiness score significantly different from the hypothesized value 5. The mean happiness score observed in the sample is 6.37265, ranging from 2.5 to 8.6. This quantitative measure suggests that, on average, respondents reported a relatively high level of happiness.We used a Wilcoxon signed-rank test. The test produced the following results:

V = 5592.5 p-value = 1.068e-11

2) **Is there a significant difference between the means of happiness scores of cities with high traffic density and medium traffic density?**

By exploratory analysis, we have seen that happiness score medians are 5.65 for high-traffic density cities and 7 for medium-traffic cities. Distributions are roughly similar as seen in the histograms. Requirements for the t-test weren't satisfied, so we used the Mann-Whitney-Wilcoxon test to determine whether there was a significant difference. With the test result, we found that population distributions and medians are different. Therefore, we can say that the happiness score means of high-traffic density cities and medium-traffic density cities are significantly different.

3) **Do countries with air quality above %60 degrees have higher happiness rates?**

This dataset shows that more cities have an air quality index below 60 compared to cities above 60. Shapiro-Wilk test is used and the result of the Shapiro-Wilk Test W is approximately 0.84205 and the p-value is associated with 0.0000008124. Because the p-value is less than 0.05, the null hypothesis is rejected. As a result, data is not normally distributed but we assume it is normally distributed. After that t-test is used and the result of the t-test, the p-value is 1 and fails to reject the null hypothesis. As a result, Countries with air quality above %60 degrees have higher happiness rates.

4) Is there a significant difference in the proportion of cities with a high healthcare index between cities with high green space areas and low green space areas?

The analysis aimed to determine whether there is a significant difference in the proportion of cities with a high healthcare index between cities with high green space area and low green space area. The mean happiness score observed in the sample is 6.37265, ranging from 2.5 to 8.6. This quantitative measure suggests that, on average, respondents reported a relatively high level of happiness.

Using a Fisher's Exact test, we compared the these two groups. The test produced the following results:

p-value< 2.2e-16 95% , Confidence Interval: [33.97613, 976.15094] , Odds Ratio: 146.7585

5) How does the Healthcare Index affect the Happiness Score?

Healthcare Index and Happiness Score is visualized by scatter plot and there is a positive linear relationship between them.Also, the equation is Happiness_Score=97.322−1.527·Healthcare_Index.The intercept is 97.322 and $\beta1$=-1.527

Also, Healtcare index and Happiness Score is visualized by histogram. They have a left skewed distribution. They need to meet certain assumptions in order to examine their contents. These assumptions are linearity, independence, homoscedasticity, normality of residuals. The result of the linearity, correlation coefficient is 0.8470444 so there is a positive strong relationship. To check independence, Durbin Watson Test is used and the independence assumption is met. The graph of homoscedasticity is a smoothly running graph. It indicates that model satisfies the homoscedasticity assumption.Q-Q plot has a straight line forms so model has a normal distribution.

What Air Quality Index and Healthcare Index affect Happiness Score?

We researched the effect of the Air Quality index and Green Space Area on happiness scores.The intercept of the coefficients is 8.883840 and the standard error is 0.241746. As a result the regression equation is Happiness Score=8.883840-0.026509(Air Quality Index)-0.005353(Green Space Area).Also, multiple R-squared is 0.7332 and adjusted R-squared 0.7285 so High values indicate that the model explains the variance of the dependent variable very well. There are some assumptions for multiple models. These are linearity, normality of residuals, homoscedasticity, multicollinearity, and independence of errors. Checked whether they were met or not. Firstly, The relationship between variables and happiness score was examined. There is a positive weak linear relationship between Green Space Area and Happiness Score because the correlation coefficient is 0.3925. There is a negative strong relationship between the air quality index because the correlation coefficient is -0.8345. After that Q-Q plot is visualized and the graph has a straight line so there is a normal distribution between the two residuals. Homoscedasticity is visualized and the distribution graph is straight and a constant line forms. It indicates that homoscedasticity is met. Multicollinearity is checked and the variation inflation factor is 1. 715366 so there is moderate multicollinearity and there is no multicollinearity problem. To check the independence of errors Durbin Watson Test is used. The result of the test, the null hypothesis is rejected. Therefore, this situation violates the independence of errors assumption. However, we assume that the independence of errors assumption is satisfied.

6) Do means of happiness scores differ depending on the healthcare index?

By exploratory analysis, we have seen that the mean happiness score is 6.37, and the mean healthcare index is 80.6. By the plots, we have seen that there was a visible difference in happiness scores between low and high healthcare indexes. Checked the requirements for ANOVA and found that data is eligible for the ANOVA test. We found that the calculated p-value is 2e-16. We rejected the null hypothesis and concluded that population means are significantly different. Comparing population means among each other, we concluded that 1st healthcare index group, which has the lowest indexes, has a significantly lower happiness score mean than 2nd and 3rd groups. The happiness score means of the 2nd and 3rd groups are not significantly different from each other.

Conclusion

1) **Is the average happiness score significantly different from the hypothesized value 5?**
   The p-value obtained from the test is significantly less than the conventional significance level of 0.05, indicating that we can reject the null hypothesis. Therefore, we conclude that the median happiness score in the population is significantly different from the hypothesized value of 5.

2) **Is there a significant difference between the means of happiness scores of cities with high traffic density and medium traffic density?**
   We found a significant difference in the means of happiness score between high-traffic density cities and medium-traffic density cities. It shows that traffic affects the happiness levels of people and traffic density of a city is one of the determining factors of happiness.

3) **Do countries with air quality above %60 degrees have higher happiness rates?**
   Research question variables visualized by box plot. It was checked whether the conditions were met and the necessary tests were performed. As a result, Countries with air quality above 60% degrees have higher happiness rates.

4) **Is there a significant difference in the proportion of cities with a high healthcare index between cities with high green space areas and low green space areas?**
   The p-value obtained from the t-test is significantly less than the conventional significance level of 0.05, indicating that we can reject the null hypothesis. Therefore, the results suggest that there is a statistically significant difference in the proportion of cities with a high healthcare index between cities with high green space area and low green space area.

5) **How does the Healthcare Index affect the Happiness Score?**
   We proved that it satisfies the assumptions for applying linear regression. As a result,equation is Happiness_Score=97.322−1.527·Healthcare_Index. We made tests according to these assumptions. As a result, there is a strong positive relationship between Healthcare Index and Happiness Score.

   What Air Quality Index and Healthcare Index affect Happiness Score?
   The necessary assumptions for multiple linear regression were examined and the necessary tests were applied. These tests were visualized with a Q-Q plot and homoscedasticity scatter plot. Green space area affects positively but air quality index affects negatively. Then, other assumptions are tested. Also, regression line equation is Happiness Score=8.883840-0.026509(Air Quality Index)-0.005353(Green Space Area).

6) **Do means of happiness scores differ depending on the healthcare index?**
   We have seen that cities with the lowest healthcare indexes have significantly lower happiness scores compared to cities with medium and high level healthcare indexes. Healthcare seems to be a determining factor of happiness. However, it does not significantly affect happiness levels when the index is higher than 75.