A Project Report

On

HEALTHY LIFESTYLE CITIES REPORT 2021

By

Group No - 3

WSU ID:	**NAME:**	**E-MAILS:**
K968X456	Vijaya Ramya CH	vxchilimutha@shockers.wichita.edu
F233Z974	Vinay Chowdari Mandava	vxmandava2@shockers.wichita.edu
Q684A655	Ravikiran Nallamothu	rxnallamothu@shockers.wichita.edu
J639H476	Lokesh Muppalla	lxmuppalla@shockers.wichita.edu

WICHITA STATE UNIVERSITY

Submitted to

Mrs. Rozsa Zaruba

CS797M - Introduction to Linear Data Modeling
School of Computing Department
Wichita State University

ABSTRACT:

After the pandemic, most people are putting the emphasis on health and food hygiene. With such a crazy lifestyle, stress at work, lack of sleep, and extensive use of gadgets, we are gradually transforming into the generation of burning-still-tired monsters.

Given this, there is no wonder why an increasing number of people decide to change their current lifestyle for the better, which usually means more physical activity, better food choices, regular meditation, and alike. Some people even move to a different city, where it would be easier for them to change their habits and integrate a healthier lifestyle.

This study examines the various factors that affect the level of happiness in a city. Many models have been developed using variables to get the best fit model to get the best results for the happiness level

INTRODUCTION:

This case study is to show the relationship between cities happiness levels and other key factors effecting. Your geographic location can have a significant bearing of many parts of your life, including your income potential, your health, and the activities you do outside of work. It wasn't easy to decide which criteria to consider. Eventually, we decided on the following: Sunshine hours, Cost of a bottle of water, Obesity levels, Life expectancy, Pollution, Happiness levels, Outdoor activities, Number of take-out places. We performed data visualization to get better instincts from the data and generated models to decided best factors effecting happiness levels of a city.

PROBLEM STATEMENT:

This research paper explains the Ranking of cities according to their healthy lifestyles. The team at Lenstore has analyzed 44 cities across the globe to uncover where it's easier to lead a well-rounded, healthy lifestyle. From obesity levels to pollution rates, each city has been scored across 10 healthy living metrics.

My team and I have chosen to predict the happiness level in a city by all other factors found in the Data set. As we know not all of the factors can affect the happiness we need to choose correct factors by suitable methods and to find the best model which describes happiness level based on those suitable Predictor variables.

PROBLEM METHODOLOGY:

- Clean the Data set to use it for the analysis.
- Choose the respective predictor variables and response variables from the data set and create a Full Model by Multiple Linear Regression.
- Finding a reduced model containing effective predictors by Variable Selection Methods which can determine the chosen Response variable.
- Comparing the Models by ANOVA and by MSE values.
- Checking for Assumptions of the Model and If necessary, transforming the Model by Box-Cox.

A. Data Description:

The chosen data set must be cleaned so that we can work on it for the processing. This is how the original data set looks like.

City	Rank	Sunshine hours(City)	Cost of a bottle of water(City)	Obesity levels(Country)	Life expectancy(years) (Country)	Pollution(Index score) (City)	Happiness levels(Country)	Outdoor activities(City)	Number of take out places(City)	Cost of a monthly gym membership(City)
Amsterdam	1	1858	£ 1.92	20.40%	81.2	30.93	7.44	422	1048	£ 34.90
Sydney	2	2636	£ 1.48	29.00%	82.1	26.86	7.22	406	1103	£ 41.66
Vienna	3	1884	£ 1.94	20.10%	81	17.33	7.29	132	1008	£ 25.74
Stockholm	4	1821	£ 1.72	20.60%	81.8	19.63	7.35	129	598	£ 37.31
Copenhagen	5	1630	£ 2.19	19.70%	79.8	21.24	7.64	154	523	£ 32.53
Helsinki	6	1662	£ 1.60	22.20%	80.4	13.08	7.8	113	309	£ 35.23

It contains data which has many datatypes in the table. It contains Strings, Serial Numbers, percentages, and currencies with their respective symbols before the numbers.
We cleaned the data so that it should be compatible for processing in finding a better model. We have removed city, Rank as our Goal is to find the Happiness level based on different factors in a City. We have shortened our column names so it would be easier to view the analysis. Here are the shortened names for Columns.

Sunshine hours(City) - Sshnehrs
Cost of a bottle of water(City) - Costofwtr
Obesity levels(Country) - Owght
Life expectancy(years) (Country) - Lfexptncy
Pollution (Index score) (City) - PollutionIS
Happiness levels(Country) - HappLevel
Outdoor activities(City) - OutAct
Number of take out places(City) - N0.ofTOplcs
Cost of a monthly gym membership(City). – CostofGym

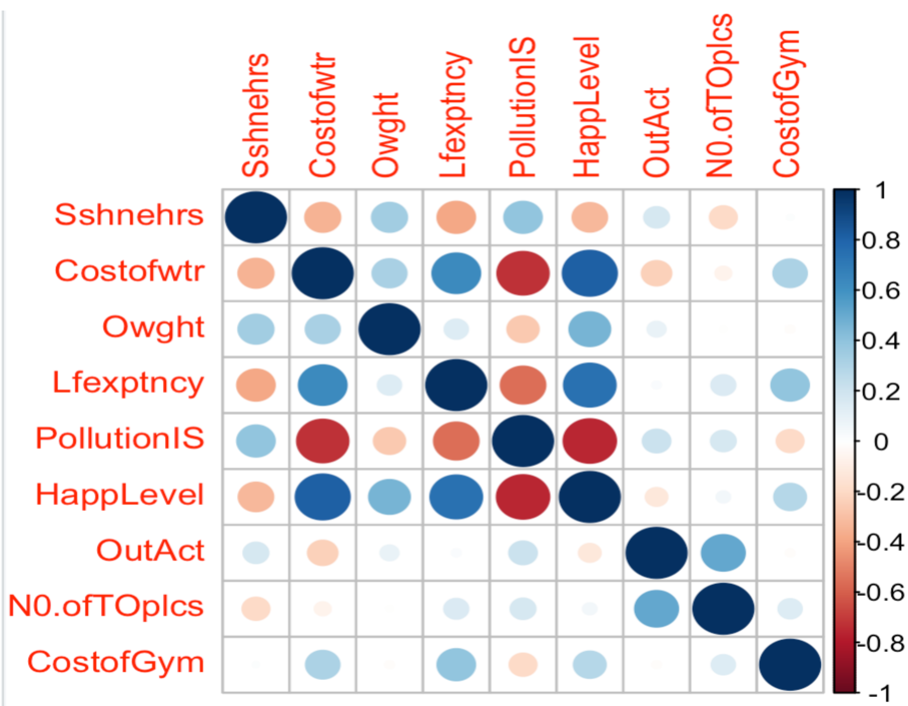
B. Data Cleaning:

Here is the picture of cleansed data which is ready for processing:

Sshnehrs	Costofwtr	Owght	Lfexptncy	PollutionIS	HappLevel	OutAct	N0.ofTOplcs	CostofGym
1858	1.92	0.20	81.2	30.93	7.44	422	1048	34.90
2636	1.48	0.29	82.1	26.86	7.22	406	1103	41.66
1884	1.94	0.20	81	17.33	7.29	132	1008	25.74
1821	1.72	0.21	81.8	19.63	7.35	129	598	37.31
1630	2.19	0.20	79.8	21.24	7.64	154	523	32.53
1662	1.60	0.22	80.4	13.08	7.8	113	309	35.23

C. Check the Correlation:

We checked correlation between all the variables to see there are related to each other.



D. Regression models:

1. Full Model(model1):

We created our Full model by choosing **HappLevel** as the response variable and remaining variables as predictor Variables and performed Multiple Linear Regression.

```
> #Building the Full model.
> model1=lm(formula=HappLevel~.,data=happ)
> summary(model1)

Call:
lm(formula = HappLevel ~ ., data = happ)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.92570 -0.18622 -0.00367  0.18808  0.93701 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  1.202e+00  1.495e+00   0.804  0.42732   
Sshnehrs    -8.994e-05  1.829e-04  -0.492  0.62612   
Costofwtr    4.025e-01  1.689e-01   2.383  0.02307 * 
Owght        2.676e+00  9.099e-01   2.941  0.00593 **
Lfexptncy    6.481e-02  1.872e-02   3.462  0.00150 **
PollutionIS -1.242e-02  5.151e-03  -2.412  0.02160 * 
OutAct      -5.453e-04  7.050e-04  -0.773  0.44473   
N0.ofTOplcs  5.899e-05  6.645e-05   0.888  0.38106   
CostofGym   -3.848e-04  5.648e-03  -0.068  0.94609   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4398 on 33 degrees of freedom
Multiple R-squared:  0.843,	Adjusted R-squared:  0.8049 
F-statistic: 22.14 on 8 and 33 DF,  p-value: 3.527e-11
```

2. Finding the Best Reduced Model:

We have tried choosing variables based on strong correlation and other methods but when we compared to the Full Model none of them seemed to be better when we performed Anova between them.

- Variable Selection by Mallow’s CP

We have considered to pick a model based upon the Lowest Mallow’s CP value and to consider those predictor variables. Here are the variables chosen for the reduced model.

```
> #variable selection by Lowest Mallow's CP value
> ols_step_all_possible(model1)
   Index N                Predictors    R-Square Adj. R-Square Mallow's Cp
2      1 1                 Costofwtr 0.655230403   0.646611163   34.450843
5      2 1               PollutionIS 0.575712568   0.565105382   51.160942
4      3 1                 Lfexptncy 0.552032946   0.540833769   56.137044
3      4 1                     Owght 0.209824725   0.190070343  128.049632
1      5 1                  Sshnehrs 0.106570355   0.084234614  149.747793
8      6 1                 CostofGym 0.078667591   0.055634281  155.611357
6      7 1                    OutAct 0.016512267  -0.008074926  168.672850
7      8 1               N0.ofTOplcs 0.002049998  -0.022898752  171.711991
17     9 2       Costofwtr Lfexptncy 0.745849640   0.732816289   17.407864
27    10 2     Lfexptncy PollutionIS 0.728326803   0.714394845   21.090162
18    11 2     Costofwtr PollutionIS 0.716162347   0.701606570   23.646435
16    12 2           Costofwtr Owght 0.700241992   0.684869786   26.991982
22    13 2           Owght Lfexptncy 0.676020454   0.659406118   32.081964
```

```
80     88 3                 Owght OutAct N0.ofTOplcs 0.265525330   0.207540488 120.344552
56     89 3              Sshnehrs OutAct CostofGym 0.191099202   0.127238613 135.984665
57     90 3          Sshnehrs N0.ofTOplcs CostofGym 0.190685025   0.126791738 136.071702
55     91 3            Sshnehrs OutAct N0.ofTOplcs 0.112832711   0.042793188 152.431804
92     92 3              OutAct N0.ofTOplcs CostofGym 0.099964351   0.028908905 155.135997
128    93 4     Costofwtr Owght Lfexptncy PollutionIS 0.833918196   0.815963407   2.900893
93     94 4       Sshnehrs Costofwtr Owght Lfexptncy 0.813319563   0.793137894   7.229547
108    95 4      Sshnehrs Owght Lfexptncy PollutionIS 0.806213454   0.785263557   8.722844
148    96 4        Owght Lfexptncy PollutionIS OutAct 0.803483279   0.782238228   9.296571
```

- Here is the summary of Reduced Model.

```
> #Best Reduced model
> model5=lm(formula=HappLevel~Costofwtr+Owght +Lfexptncy+PollutionIS,data=happ)
> a=summary(model5)
> a

Call:
lm(formula = HappLevel ~ Costofwtr + Owght + Lfexptncy + PollutionIS,
    data = happ)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06976 -0.16953  0.00274  0.15453  0.89252

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.700013   1.320600   0.530 0.599229
Costofwtr    0.438897   0.155558   2.821 0.007645 **
Owght        2.325946   0.706123   3.294 0.002182 **
Lfexptncy    0.068491   0.016402   4.176 0.000173 ***
PollutionIS -0.012395   0.004545  -2.727 0.009705 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4272 on 37 degrees of freedom
Multiple R-squared:  0.8339,    Adjusted R-squared:  0.816
F-statistic: 46.45 on 4 and 37 DF,  p-value: 6.188e-14
```

We compared R-Squared values, F -Statistic values and Residual Standard Error between this and Full Model. We compared this reduced model to the full model by performing Anova and found that this is best model.

3. ANOVA Model:

```
> #Comparing Models By performing Anova.
> anova(model5,model1)
Analysis of Variance Table

Model 1: HappLevel ~ Costofwtr + Owght + Lfexptncy + PollutionIS
Model 2: HappLevel ~ Sshnehrs + Costofwtr + Owght + Lfexptncy + PollutionIS +
    OutAct + N0.ofTOplcs + CostofGym
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     37 6.7520
2     33 6.3842  4   0.36775 0.4752 0.7536
```

- By the above results we can say that p-value for Full Model is very high, and we cannot reject the Null Hypothesis. So Reduced model is better than Full Model.

E. Check the VIF:

We checked weather there was any presence of Multicollinearity between the predictor variables with the help of VIF Factor, there isn't any as all of them has VIF less than 10.

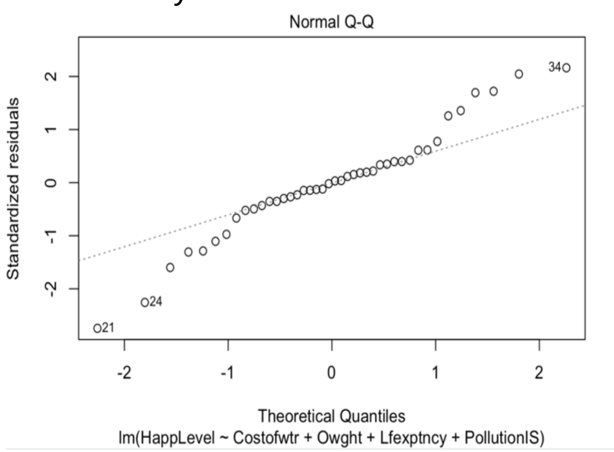
```
#Checking for Multicollinearity.
vif(model1)
 Sshnehrs   Costofwtr       Owght   Lfexptncy PollutionIS      OutAct N0.ofTOplcs   CostofGym
 2.289383    2.939441    1.758554    2.105895    2.671342    1.622980    1.849409    1.405279
```

ANALYSIS AND RESULTS:

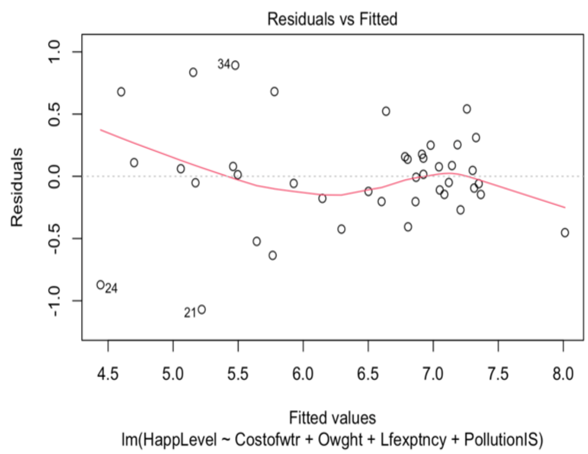
I. Checking for Assumptions:

When checked for assumptions the plots for Linearity and Normality, they do not hold. Here are the plots below:

Normality:



Linearity:



As normality do not hold, we have performed BOX-COX Transformation to the above model and here are the results.

```
> library(MASS)
> #Transforming the Model[Box-cox Transformation].
> bc=boxcox(happ$HappLevel~happ$Costofwtr+happ$Owght +happ$Lfexptncy+happ$PollutionIS, data = happ)
> lambda=bc$x[which.max(bc$y)]
> lambda
[1] 2
> BXCX <- lm(((happ$HappLevel^lambda-1)/lambda) ~happ$Costofwtr+happ$Owght +happ$Lfexptncy+happ$PollutionIS,data = happ)
> b=summary(BXCX)
> b

Call:
lm(formula = ((happ$HappLevel^lambda - 1)/lambda) ~ happ$Costofwtr +
    happ$Owght + happ$Lfexptncy + happ$PollutionIS, data = happ)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0580 -1.2959 -0.3027  0.9280  5.0195

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -9.36795    7.49478  -1.250 0.219173
happ$Costofwtr    2.87712    0.88283   3.259 0.002401 **
happ$Owght       13.31924    4.00745   3.324 0.002011 **
happ$Lfexptncy    0.35826    0.09308   3.849 0.000454 ***
happ$PollutionIS -0.08191    0.02579  -3.176 0.003011 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.424 on 37 degrees of freedom
Multiple R-squared:  0.8503,    Adjusted R-squared:  0.8341
F-statistic: 52.53 on 4 and 37 DF,  p-value: 9.254e-15
```

After the transformation, model seemed to be improved than the previous model and even Normality also Improved.

II. Comparing MSE's:

```
> #MSE Value of Final Reduced Model
> mean(a$residuals^2)
[1] 0.1607608
> #MSE Value of Transformed Model
> mean(b$residuals^2)
[1] 5.17792
```

But when we compared the MSE's of Reduced Model and Transformed Model Reduced model's MSE seemed too less. So, we stick to the reduced model.

But when coming to the linearity of the model, there was curve in Residuals Vs Fitted plot. So, we need to transform one of the predictor variables by applying square to them to fix Linearity assumption.

Here are the combinations we tried to fix linearity.

So, we stick with the Final reduced model which can predict the Happiness levels of the People based upon Cost of a bottle of water, Obesity levels, Life expectancy and Pollution (Index score) of the City.

- ➔ We tried different combinations by applying square to different predictor variables, but linearity doesn't hold very well.

```
#Trying to fix Residual vs Fitted Graph.

#2nd order Multilinear Regression
model51=lm(formula=HappLevel~Costofwtr^2+Owght +Lfexptncy+PollutionIS,data=happ)
plot(model51)
model52=lm(formula=HappLevel~Costofwtr+Owght^2 +Lfexptncy+PollutionIS,data=happ)
plot(model52)
model53=lm(formula=HappLevel~Costofwtr+Owght +Lfexptncy^2+PollutionIS,data=happ)
plot(model53)
model54=lm(formula=HappLevel~Costofwtr+Owght +Lfexptncy+PollutionIS^2,data=happ)
plot(model54)

#2nd Order Interaction Multilinear Regression
model56=lm(formula=HappLevel~Costofwtr*Owght+Owght*Lfexptncy +Lfexptncy^2+PollutionIS*Costofwtr,data=happ)
plot(model56)
#3rd Order Multilinear Regression
model55=lm(formula=HappLevel~Costofwtr^3+Owght^3 +Lfexptncy^3+PollutionIS^3,data=happ)
plot(model55)
```

CONCLUSION:

The purpose of this research was to identify an effective model with significant variables to calculate the happiness levels of cities. Based on the regressions conducted, we can conclude that there is the possibility of developing multiple reduced models to compare and conclude to best fit model. Comparing models, we declared variables cost of a water bottle, obesity levels, life expectancy, and pollution index score are significant variables to determine happiness levels. Further transformation of the model using box cox transformation did not help the model improve. Our strategy for the future is to collect additional data and train the model on it to improve accuracy.

By statistical analysis from the chosen Data set, Happiness level in each city mostly depends upon the factors like Cost of a bottle of water, Obesity levels, Life expectancy, Pollution (Index score) with 83.40% accuracy.

LESSONS THAT WHAT WE HAVE LEARNED:

a. We need to follow the statistical methods and results for selecting the predictors not on logical factors, which doesn't work all the time. It also affects the goodness of the Reduced Model.
b. Transforming the model by Box-Cox isn't always makes a model better.
c. We need to give priority to MSE than Linearity assumption of the Model.
d. Variable selection methods are very effective in considering correct predictors for the reduced model when there are many variables to choose from.
e. Variance increases with the number of predictors as we observed from Full Model and reduced models in Multi Linear Regression.

BIOGRAPHY:

❖ Name: Vijaya Ramya CH
Major: Computer Science
Future Aim: Working with datasets always excited me. WSU helped me with hands-on practice with projects to face real-world issues.

❖ Name: Vinay Chowdari Mandava
Major: Data Science
Future Aim: Initially, I am an undergraduate student from ECE. I'm interested in integrating visualization, programming, and statistics to create more clear information about data, so I decided to learn more about it. Now that I'm at WSU, I'm on the way to reach my goal.

❖ Name: Ravikiran Nallamothu
Major: Data Science
Future Aim: Today's world is generating a lot of data every second. My goal as Data Scientist is to explore, sort and analyze that mega data from various sources to take advantage of them and reach conclusions to optimize business processes or for decision support.

❖ Name: Lokesh Muppalla
Major: Data Science
Future Aim: I've been fascinated by data science since I finished my bachelor's degree. I believe my dream will come true because I am approaching to my goal.

APPENDIX:

- https://www.gfmag.com/global-data/non-economic-data/best-cities-to-live
- https://ourworldindata.org/obesity
- http://happyplanetindex.org/countries
- https://en.wikipedia.org/wiki/List_of_cities_by_sunshine_duration
- https://www.numbeo.com/pollution/rankings.jsp
- https://worldhappiness.report
- https://www.numbeo.com/cost-of-living
- https://worldpopulationreview.com/country-rankings/average-work-week-by-country
- https://data.oecd.org/emp/hours-worked.htm
- https://www.tripadvisor.co.uk