Northeastern University
College *of* Engineering

STATISTICAL METHODS IN ENGINEERING
IE-7280

PROJECT REPORT

1. Modeling Average Rent Prices Across States in the U.S.A. using Regression Analysis
2. Evaluating Relationship Between Economic Progress and Climate Change Across Countries Using 2-Factor ANOVA

By
Sahar Tariq
12/19/2020

Contents

MODELING AVERAGE RENT PRICES ACROSS STATES IN THE U.S.A USING REGRESSION ANALYSIS 3

1. ABSTRACT 3

1.1 Introduction: 3

1.2 Objective Statement: 3

1.3 Statistical Procedure: 3

1.4 Data Description: 3

2. DATA PREPROCESSING 4

3. DATA ANALYSIS & DISCUSSION 5

3.1 Model Building: Simple Linear Regression 5

3.2 Model Building: Adding Variables to Model - Multiple Linear Regression 7

3.3 Finetuning Model: Multiple Linear Regression 8

4. DATA EVALUATION 9

5. CONCLUSION 9

EVALUATING RELATIONSHIP BETWEEN ECONOMIC PROGRESS AND CLIMATE CHANGE ACROSS COUNTRIES USING 2-FACTOR ANOVA 10

1. ABSTRACT 10

1.1 Introduction: 10

1.2 Objective Statement: 10

1.3 Statistical Procedure: 10

1.4 Data Description: 10

2. DATA PREPROCESSING 11

3. DATA ANALYSIS & DISCUSSION 12

3.1 Significance of Effect Analysis: 2-Factor ANOVA 12

3.2 Discussion of Results: Effects and Interaction Plots 14

4. DATA EVALUATION 15

5. CONCLUSION 16

DATA SOURCES 17

REFERENCES 17

MODELING AVERAGE RENT PRICES ACROSS STATES IN THE U.S.A USING REGRESSION ANALYSIS

1. ABSTRACT

1.1 Introduction:

The median rent in the U.S. has steadily increased in the last decade. Americans are spending about 37% of their income on rent, leaving inadequate finances for essential expenses, higher education, and retirement. [1] Citizens are flocking to metropolitans in the pursuit of a better quality of life, as evidenced by the exponential growth in their population densities. However, with higher earning potential comes extreme rent prices, raising the question: Are high-income locations worth it?

This report statistically analyzes if higher the income in a state, higher the rent, keeping affordability the same in each state regardless of income. Furthermore, this report models what core socioeconomic variables are directly affecting the rent across America's states.

1.2 Objective Statement:

The objective is to test the hypothesis if Average Rent of a 1-Bedroom House is affected linearly by Median Household Income. A second objective is to create a model for Average Rent of a 1-Bedroom House vs socioeconomic variables to find which factors are responsible for the rent prices across America's states.

1.3 Statistical Procedure:

A simple linear regression model of Median Household Income (predictor variable) on Average Rent of 1-Bedroom House (response variable), along with regression analysis and hypothesis testing will answer if there is linearity between the two variables.

Several multiple linear regression models of different socioeconomic variables will be compared using sequential model selection methods to find the best fit model for Rent. Statistical computations will be done in R programming language.

1.4 Data Description:

This analysis was done using 2016 data for the 50 U.S. states.

The trendline for Median Income over time per state from 1984 to 2019 (in today's dollars), and trendline for Average Rent over time per state from 2010-2020 have maintained a steady average positive gradient and uniform variation between the states. Therefore, it is adequate to use a point in time. Our analysis is focused on the difference in rent and income between the different states. 2016 was found to be a year without much external influences from factors such as recession, pandemic, war.

The variables from the different sources were combined to create one relational table, grouped by state. The first row of the 50x10 table is shown below.

	State <chr>	Population <int>	Density <dbl>	GDP <int>	GDP_Capita <int>	MinWage <dbl>	Bachelors <dbl>	Crime <dbl>	Income <int>	Rent <int>
1	Alabama	4860545	96.8	207608	42689	7.25	27.2	532	47221	709

Figure 1: Dataset structure

[1] https://www.jchs.harvard.edu/sites/jchs.harvard.edu/files/05_harvard_jchs_americas_rental_housing_2017.pdf

The following table lists the data sources and data characteristics of the variables used:

	Table 1: Data Description		
	Variable	**Data Source**	**Data Description**
1	Average Rent of 1-Bedroom House ($)	Zillow - an American online real estate database company	Mean cost of 1-bedroom rental houses that fall into the 40th to 60th percentile range, on 1/31/2016, by state
2	Median Household Income ($)	United States Census Bureau	Income in 2016 dollars, by state
3	Population Density (per sq mile)	United States Census Bureau	Average number of people living per square-mile in 2016, by state
4	GDP (millions of $)	Bureau of Economic Analysis- U.S. Department of Commerce	Gross Domestic Product, in 4th quarter of 2016, by state
5	GDP per Capita ($)	Calculated	GDP divided by population, per state
6	Minimum Wage ($)	U.S. Department of Labor	Required minimum wage, in 2016, by state
7	Percent of population with a Bachelors degree (%)	National Science Foundation	Percent of population having a Bachelor's degree, living in each state. Degree not necessarily acquired from this state.
8	Crime Rate (per 10,000 People)	FBI UCR	Violent crime rate per 100,000 people, by state. Violent crime includes offenses of murder, rape, robbery, and aggravated assault.

2. DATA PREPROCESSING

Data Preprocessing steps taken:

- Data cleaning: Removed characters such as, "" $ % ~ from the datasets
- Subsetting: Only kept 2016 data for the 50 states
- Combining: State-level only, then left-outer joined to combine the different data sets into a one table
- Checking for outliers: The boxplots below show there are no unreasonable outliers. The few outliers cannot be removed as they are for states with very high/low values.

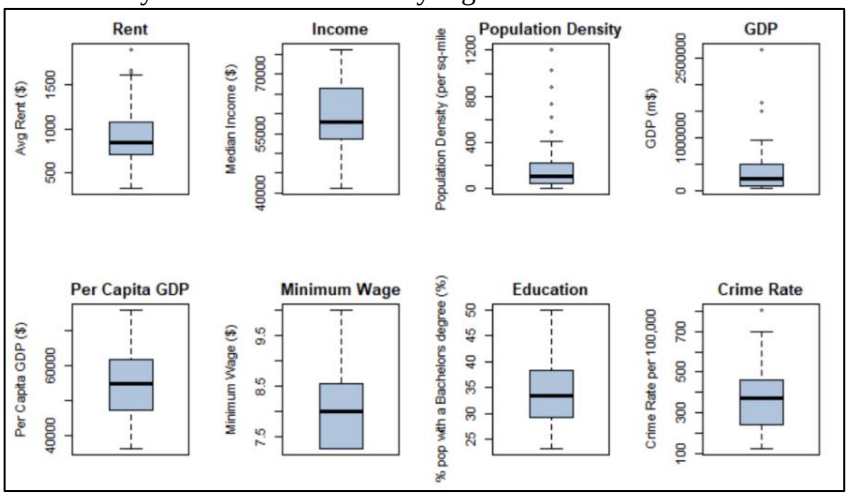

Figure 2: Boxplots of Variables

- Checking if data transformation is required: The distribution plots below show

- Income, Per Capita GDP, Education % and Crime have bell-shaped distribution and nonsignificant skewness, meaning they are normally distributed.
- Rent has slightly left leaning shaped distribution and some skewness, meaning it is slightly deviated from normal distribution but not enough to require any transformation.
- Population Density and GDP are significantly left skewed. Log, square, and square root transformations were attempted, however that did not significantly improve the skewness or model, so transformation will not be applied.

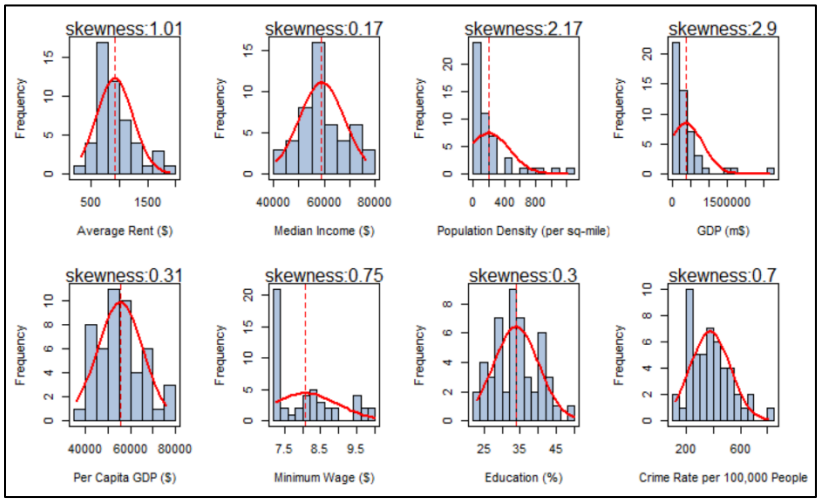

Figure 3: Distribution of Variables

3. DATA ANALYSIS & DISCUSSION

3.1 Model Building: Simple Linear Regression

The claim is there is linearly proportional relationship between Average Rent and Median Household Income. This means higher the income, higher the rent in a state, making a lower income state just as good to live in as a higher income state, if rent-affordability was the sole indicator for quality of living.

The scatterplot's trendline has correlation coefficient **R=0.5815097**.
R values over 0.5 indicate there is some linear relationship.

Correlation coefficient is derived by:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Figure 4: Relationship Between Rent and Income

The linear model relating Income and Rent created gives the regression statistics:

```
Call:
lm(formula = df_1$Rent ~ df_1$Income)

Coefficients:
(Intercept)  df_1$Income
 -308.57030      0.02098


Call:
lm(formula = df_1$Rent ~ df_1$Income)

Residuals:
    Min      1Q  Median      3Q     Max
-571.97 -169.62  -34.34  185.95  690.13

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.086e+02  2.526e+02  -1.222    0.228
df_1$Income  2.098e-02  4.237e-03   4.952 9.48e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 266.8 on 48 degrees of freedom
Multiple R-squared:  0.3382,    Adjusted R-squared:  0.3244
F-statistic: 24.52 on 1 and 48 DF,  p-value: 9.485e-06

[1] 689.0923
[1] 698.6524
```

The model has $\boldsymbol{\beta_1 = 0.02098}$. The basis of model linearity and adequacy is testing if β_1 is not 0. **If β_1 is 0, there is essentially no relationship between Rent and Income**. The table below shows how the model statistics were used in hypothesis testing methods to test the claim that there is linearity between Rent and Income, at a α=0.05 level of significance.

Table 2: Analysis of Simple Linear Regression Model

Statistic	Purpose of Statistic	Analysis Criteria	Conclusion
Multiple R^2	Variance proportion: Shows the proportion of variance in Rent accounted for by Income		Model R^2 =0.3382 33.82% of variance in Rent due to changes in Income
F-statistic	Lack of fit test: Checks if β coefficients are 0. If coefficients are 0, the model has a lack of fit	Ho: β1 = 0 (significant lack of fit) Ha: β1 not 0 Reject Ho if $F>F_{0.05,k,n-2}$	Model F=24.52 $F_{\alpha,1,48}$=4.04 Reject Ho. There is no significant lack of fit. → Linear model fits data adequately
t-value	Linearity test: Checks if the value of ρ is 0. If ρ is 0, β_1 = 0, meaning there essentially is no linear regression	Ho: ρ = 0 (no linearity) Ha: ρ not 0 Reject Ho if $t>t_{0.05,n-2}$	Model t =4.952 $t_{\alpha,48}$=1.68 Reject Ho. There is not enough proof of lack of linearity → Variables have linear relation
Pr(>|t|)	Linear model adequacy test	Ho: β1=0 (linear model is not adequate) Ha: β1 not 0 Reject Ho if p<0.05	Model $p=9.484x10^{-6}$ Reject Ho. There is not enough proof of model inadequacy → Model is adequate

At a 0.05 level of significance, it cannot be disproved that there is linear relationship between Rent and Income. Therefore, we can conclude that there is linear relationship, implying higher income states have linearly higher rent. Therefore, disposable income after rent remains the same throughout all U.S. states → Living in a higher income state and earning higher income does not make people richer.

3.2 Model Building: Adding Variables to Model - Multiple Linear Regression

Since only 33.82% of the variation in Rent is accounted for by Income, it is evident there are other variables affecting the average rent in states. Research shows the top socioeconomic factors that can affect rent are Population Density, GDP, Productivity per capita, Education level, Crime, and Minimum Wage. To avoid overfitting due to multicollinearity, predictor variables which are highly correlated are removed. Per the predictor variables correlation matrix below, variables GDP Per Capita and Minimum Wage will not be included in the regression as they are accounted for by other variables.

	Income	Density	GDP	GDP_Capita	Bachelors	Crime	MinWage
Income	1.0000000	0.3667493	0.14396919	0.7355928	0.7532571	-0.23696931	0.44080134
Density	0.3667493	1.0000000	0.20723691	0.4518667	0.5828068	-0.14923097	0.37505973
GDP	0.1439692	0.2072369	1.00000000	0.3163566	0.2125477	0.07967709	0.19752389
GDP_Capita	0.7355928	0.4518667	0.31635658	1.0000000	0.6891835	-0.09823260	0.38821426
Bachelors	0.7532571	0.5828068	0.21254768	0.6891835	1.0000000	-0.47570364	0.40761877
Crime	-0.2369693	-0.1492310	0.07967709	-0.0982326	-0.4757036	1.00000000	0.03458615
MinWage	0.4408013	0.3750597	0.19752389	0.3882143	0.4076188	0.03458615	1.00000000

Figure 5: Predictor Variables Correlation Matrix

The four chosen predictor variables, in addition to Income, are Population Density, GDP, % of Population with Bachelors Degree and Crime Rate, with correlation coefficients **R= 0.554, 0.419, 0.623 and -0.234** respectively, showing there is some moderate linear relation.

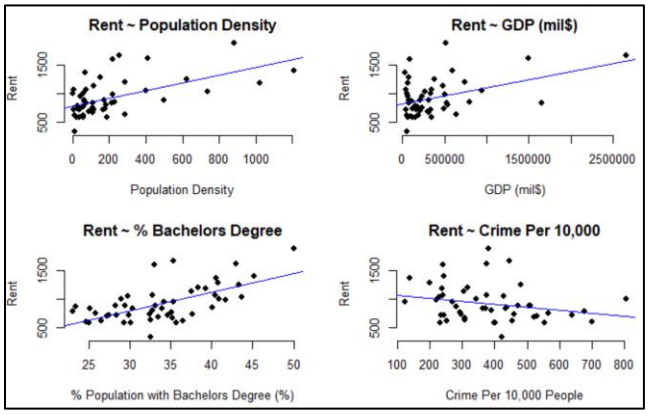

Figure 6: Relationship Between Rent and Added Predictor Variables

The multiple linear regression model created gives the regression statistics:

```
Call:
lm(formula = df_1$Rent ~ df_1$Income + df_1$Density + df_1$GDP +
    df_1$Bachelors + df_1$Crime)

Coefficients:
   (Intercept)     df_1$Income     df_1$Density        df_1$GDP  df_1$Bachelors
    -6.8644167       0.0126222       0.3763736       0.0002048       3.6001473
    df_1$Crime
    -0.2205530


Call:
lm(formula = df_1$Rent ~ df_1$Income + df_1$Density + df_1$GDP +
    df_1$Bachelors + df_1$Crime)

Residuals:
    Min      1Q  Median      3Q     Max
-431.99 -136.38  -19.75  114.55  545.66

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6.864e+00  3.004e+02  -0.023  0.98187
df_1$Income     1.262e-02  5.677e-03   2.223  0.03137 *
df_1$Density    3.764e-01  1.543e-01   2.440  0.01879 *
df_1$GDP        2.047e-04  7.185e-05   2.850  0.00664 **
df_1$Bachelors  3.600e+00  1.068e+01   0.337  0.73775
df_1$Crime     -2.206e-01  2.664e-01  -0.828  0.41211
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 225.7 on 44 degrees of freedom
Multiple R-squared:  0.5658,    Adjusted R-squared:  0.5164
F-statistic: 11.47 on 5 and 44 DF,  p-value: 4.04e-07
```

The model was tested for linearity per the hypothesis testing methodology outlined in Table 2:

- Adjusted R-squared 0.5164 shows 51.64% of variance in Rent is covered by the predictor variables
- F statistic 11.47 is more than Fa 4.04 → at least one B is not zero → **model fits data**
- t-value is more than t_α 1.68 → ρ not 0 → **model has linearity**
- P-value 4.04x10^-7 is less than 0.05 → β1 not 0 → **model is adequate**

3.3 Finetuning Model: Multiple Linear Regression

To further finetune the model, must ensure it adequately compromises between excessive bias incurred from underfitting (too few model terms) and excessive prediction variance produced from overfitting (redundancies in the model). Looking at the model's variables' coefficients, if the **p value is more than 0.05** it means variable is not significant to the linear model so should be dropped.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.864e+00  3.004e+02  -0.023  0.98187
df_1$Income    1.262e-02  5.677e-03   2.223  0.03137 *
df_1$Density   3.764e-01  1.543e-01   2.440  0.01879 *
df_1$GDP       2.047e-04  7.185e-05   2.850  0.00664 **
df_1$Bachelors 3.600e+00  1.068e+01   0.337  0.73775
df_1$Crime    -2.206e-01  2.664e-01  -0.828  0.41211
```

Figure 7: Multiple Regression Model's coefficients

To verify this, models with different combination permutations of the five variables were evaluated using sequential model selection methods.

Table 3: Sequential Model Selection

Model		AIC	BIC	PRESS	CP	adjr2	F-Statistic	p-value
	Purpose of Statistic:	Estimates relative quality of statistical models	Posterior probability of a model being true model	Measure of how well model predicts response values that were not used in building the models	Penalizes if model has non-contributing variables	% of variation in dependent variable explained by the predictor variables that actually effect it	Defines the collective effect of all predictor variables on the response variable	Goodness of fit
	Selection Criteria:	lowest	lowest	lowest	close to actual number of vars	higher	higher	good fit if p<0.05
lm(Rent ~ Income +Density +GDP+Bachelors+Crime)		691.43	704.82	3163343.43	6	0.52	11.47	0.0000004
lm(Rent ~ Income+Density +GDP+Bachelors)		690.21	701.68	3013402.96	4.69	0.52	14.26	0.0000001
lm(Rent ~ Income+Density +GDP)		689.09	698.65	2833448.37	3.48	0.52	18.83	0
lm(Rent ~ Income+Density)		695.2	702.84	3116112.18	9.49	0.45	21.02	0.0000003
lm(Rent ~ Income)		704.51	710.24	3710279.17	21.06	0.32	24.52	0.0000095
lm(Rent ~ Income+GDP)		697	704.65	3233211.67	11.46	0.43	19.44	0.0000007
lm(Rent ~ Income+Bachelors)		700.13	707.77	3488944.47	15.03	0.39	16.84	0.0000031
lm(Rent ~ Income+Crime)		705.69	713.34	3738901.45	21.98	0.32	12.59	0.0000419
lm(Rent ~ Income+Density +Bachelors)		695.96	705.52	3215549.22	10.18	0.45	14.44	0.0000009
lm(Rent ~ Income+Density +Crime)		696.61	706.17	3167172.18	10.87	0.44	14.05	0.0000012
lm(Rent ~ Income+GDP+Bachelors)		693.9	703.46	3203438.91	8.08	0.47	15.69	0.0000004
lm(Rent ~ Income+GDP+Crime)		697.06	706.62	3171218.37	11.35	0.44	13.79	0.0000015
lm(Rent ~ Income+Density +GDP+Crime)		689.56	701.03	2854727.3	4.11	0.53	14.59	0.0000001
lm(Rent ~ Income+Density +Bachelors+Crime)		697.9	709.37	3395135.19	12.12	0.44	10.62	0.0000038
lm(Rent ~ Income+GDP +Bachelors+Crime)		695.78	707.25	3403891.97	9.95	0.46	11.57	0.0000015

The top 2 multiple linear regression models are:

1) Rent=-111.9+0.0150Income+0.4170Density+0.0002GDP (model3)
2) Rent=44.7226+0.0140Income+0.4029Density+0.0002091GDP-0.2670Crime (model13)

4. DATA EVALUATION

The chosen two models' errors are evaluated using the actual data vs predicted data, to find the model with highest accuracy. [2]

Table 4: Model Error Comparison

Model		SSE	SSR	s2	MSE	RSE	RMSE	MAE	MAPE	Chi-squared statistic
	Purpose of Statistic:	Sum of difference between actual and predicted values	Sum of differences between predicted values and mean of actual variable	Variance of errors	Measure of how far from the regression line predicted data points are	Average amount response deviates from true regression	Stdev of how far from regression line data points are	Mean of absolute errors: Compare predicted and actual	Mean percent absolute errors: Forecast model errors	Measure of difference between the actual and predicted frequencies
	Selection Criteria:	lowest	lowest	lowest	lowest	lowest	lowest	lowest	lowest	lowest
	Formula:	$\sum(Y_i-\hat{Y}_i)^2$	$\sum(Y_i-\bar{Y})^2$	$\frac{SSE}{n-k-1}$	$\frac{1}{n}\sum(Y_i-\hat{Y}_i)^2$		$\sqrt{\frac{\sum(Y_i-\hat{Y}_i)^2}{n}}$	$\frac{1}{n}\sum_{-\hat{Y}_i\|}\|Y_i$	$\frac{1}{n}\sum\frac{\|Y_i-\hat{Y}_i\|}{Y_i}$	$\sum\frac{(Y_i-\hat{Y}_i)}{\hat{Y}_i}$
lm(Rent ~ Income+Density +GDP)		2317483	2845139	50380	48280	224.46	215.3	167.07	0.1982	2535
lm(Rent ~ Income+Density +GDP+Crime)		2247630	2915001	49947	46825	223.49	212.0	166.13	0.1984	2377

Selection reasoning:

- In Table 3, although Model 13 may have the 2nd best sequential model AIC, BIC, and PRESS and Model 1 has the best, the difference between the two is not too big.
- Model 13's Cp estimated the number of variables closer to the actual number of variables
- Model 13 has a better adjusted R^2 value indicating it accounts for more of the variance in Rent.
- In Table 4, Model 13 has lower errors when comparing actual data to model data

The chosen model for Rent is Model 13. It is a function of Income, Density, GDP, and Crime, with parameters Rent=44.7226+0.0140Income+0.4029Density+0.0002091GDP-0.2670Crime (model13)

5. CONCLUSION

The simple linear model affirmed the hypothesis that living in higher income states does not make one richer since mean rent in the USA is linearly proportional to median state income. Higher income potential means higher rent, leaving lesser disposable income. The multiple linear equation model *Rent=44.7226+0.0140Income+0.4029Density+0.0002091GDP-0.2670Crime* shows that mean rent is dependent on median income, population density, state GDP, and slightly on crime. This shows that perks of paying higher rent are living in a higher developed state (higher GDP) and slightly lesser crime, however it will be a more crowded state (higher population density) and similar disposable income.

[2] https://rstudio-pubs-static.s3.amazonaws.com/348670_ad694ac3e3cc475db019daa584ba663e.html

EVALUATING RELATIONSHIP BETWEEN ECONOMIC PROGRESS AND CLIMATE CHANGE IN COUNTRIES USING 2-FACTOR ANOVA

1. ABSTRACT

1.1 Introduction:

Global climate change is one of the defining issues of our world today. We have already observed shifting weather patterns effecting crop yield, increasing droughts, increasing sea levels leading to catastrophic flooding, and increasing climate refugees both in the human world and animal kingdom, to name a few. The U.N.'s hopes to limit temperature increase to maximum 1.5 °C, to avoid irreversible damages to the planet. The biggest contributor of global climate change has been increasing greenhouse gas (GHG) concentration, with CO_2 accounting for two-thirds of GHGs, mainly produced by burning fossil fuels. [3]

Despite these concerns, countries have ploughed forward in increasing production and fossil fuel consumption, in the pursuit of economic progress. Climate change in the future has become a price the present argues is worthwhile to achieve wealth. This brings us to the question, is it really? Have countries who have been increasing their CO_2 emissions or suffering from increasing temperatures been enjoying higher GDP as a reward?

This report statistically analyzes the effect of low, medium and high increase in per capita CO_2 emissions on countries' GDPs, and the effect of low, medium and high increase in average yearly temperature on a countries' GDP.

1.2 Objective Statement:

The objective is to test the claim that there is significant difference in GDP increase between High, Medium and Low levels of CO_2 increase, and between High, Medium and Low levels of temperature increase.

1.3 Statistical Procedure:

A 2-Way Analysis of Variance with Average Rate of Temperature Change (°C) as Factor 1, Average Rate of CO_2 Change (metric tons per capita) as Factor 2, and Average Rate of GDP Change (%) as the observation, will answer if the factors have significant effect on the response.

Statistical computations will be done in R programming language and Minitab.

1.4 Data Description:

This analysis was done using 1991-2016 data for 190 countries.

The GDP, average yearly temperature, and average CO_2 emissions are different between countries, due to their geographic, climate, socioeconomic, political, technological, historic differences. Therefore, the rate of change of each variable has been calculated, instead of using magnitudes. Our analysis is focused on the effect of the rate of change of temperature and CO_2 on the rate of change of GDP. Also, CO_2 is in metric tons per capita, so we will be analyzing the increase in how many metric tons of CO_2 a person of that country has been producing per year.

[3] https://www.un.org/en/sections/issues-depth/climate-change/

The variables from the different sources were combined to create one relational table, grouped by country. The first row of the 190x7 table is shown below.

1	Country_Code	Country	Temp_Change	CO2_Change	Temp_Factor	CO2_Factor	GDP_Change
2	AFG	Afghanistan	0.0647631	0.002469515	H	M	7.539524704

Figure 1: Dataset structure

The following table lists the data sources and data characteristics of the variables used:

	Table 1: Data Description		
	Variable	**Data Source**	**Data Description**
1	Rate of Change in Average Temperature (Δ°C)	The World Bank (https://climateknowledgeportal.worldbank.org/download-data)	Raw data is temperature per month for 1991 to 2016 by each country. Compute the yearly average, then find the Δ per year, then find the average Δ for each country.
2	Rate of Change in GDP (%)	The World Bank (https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?end=2016&start=1991&view=chart)	Raw data is GDP per year for 1991 to 2016 by each country. Find the Δ per year, then find the average Δ for each country.
3	Rate of Change in CO_2 emissions (Δmetric tons per capita)	The World Bank https://datatopics.worldbank.org/world-development-indicators/themes/environment.html	Raw data is emission per year for 1991 to 2016 by each country. Find the Δ per year, then find the average Δ for each country.

Note: Unable to obtain GDP and CO_2 emissions data for Kosovo, Faroe Islands, North Korea, and Somalia, and there for these countries will be excluded for this analysis.

2. DATA PREPROCESSING

Data Preprocessing steps taken:

- Removing outliers: The distribution plots in Figure 2 below show
  - Variable Rate of Change in Temperature has bell-shaped distribution and nonsignificant skewness, meaning it is normally distributed.
  - Rate of Change in CO_2 is slightly left leaning with some skewness due to some countries with very high CO_2 emissions increase but these outliers will not be removed
  - Rate of Change in GDP has quite significant skewness due to South Sudan's steeply decreasing GDP as a result of war, and Equatorial Guinea's amazingly increasing GDP as a result of striking oil in 1995.[4] Removing these two outliers improves GDP skewness to 0.72.

- Data Transformations: Rate of Temperature Change and Rate of CO_2 Change needs to be converted into factors with 3 levels, Low Medium High. This is done by partitioning along the percentiles. The boxplots in Figure 3 below show the distribution of data. The dotted horizontal lines show that the 0-25th percentile of data is categorized as Low, the 25th-75th percentile of data is categorized as Medium, and the 75th-100th percentile of data is categorized as High.

[4] https://www.bbc.com/news/world-africa-13317174

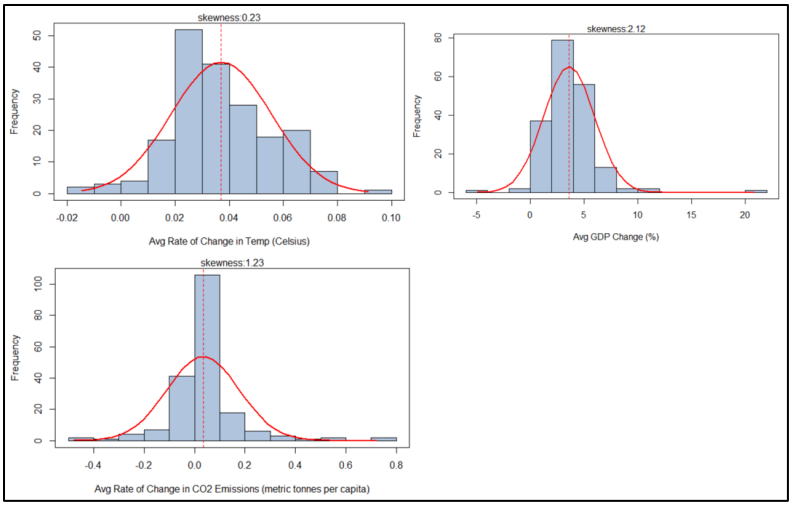

Figure 2: Distribution of Variables

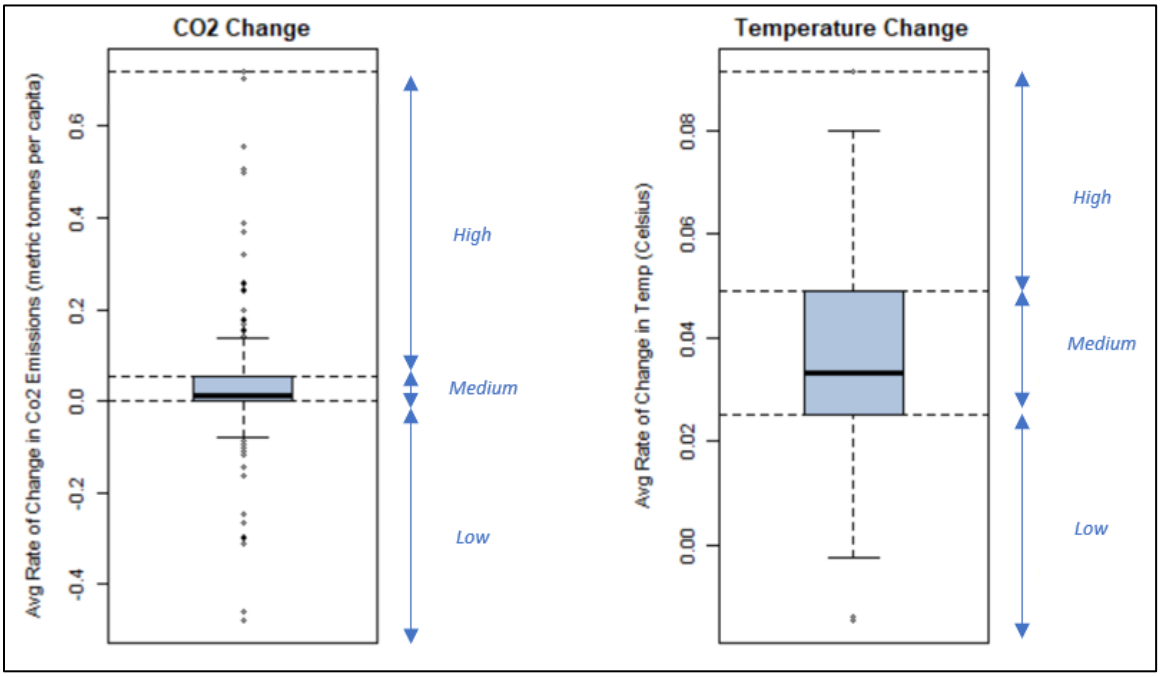

Figure 3: Categorization of Variables

3. DATA ANALYSIS & DISCUSSION

3.1 Significance of Effect Analysis: 2-Factor ANOVA

Table 3 shows how the how the data was categorized. Magnitudes of Rate of Temperature Change and Rate of CO_2 Change were partitioned in 3 levels each.

Table 2: Rate of GDP Change (%), With Rate of Temperature Change and Rate of CO_2 Change as Factors			
	Change in Temp		
Change in CO2 level	**Low j=1**	**Medium j=2**	**High j=3**
Low i=1	1.230709 0.85099 2.158963 5.342978 1.14741 3.29056 2.188402	1.830372 0.850836 4.263372 5.356744 1.292147 6.013751 1.56403 2.376436 1.279314 5.128291 1.511856 2.654289 1.709964 2.17091 2.020509 2.486108 1.746248 1.065958 1.540543 2.329254 2.07078 2.930355 3.704819 1.55568	2.294635 2.012087 4.447443 2.482288 1.885753 1.424893 5.063252 2.606466 3.013883 0.89578 1.971427 0.698619 5.04278 3.63902 4.159475 2.037767 3.670244
Medum i=2	3.224491 5.370035 3.085896 2.412256 5.435288 9.002527 4.153099 7.056538 1.108916 4.168698 2.521956 4.406468 4.413357 4.659653 5.309628 1.721472 3.194859 1.324867 2.858963 3.307053 0.906284 2.497863 0.970522 5.164858 5.908445 3.26657 1.957137 2.626211 1.912635	4.911339 1.115576 -1.80906 5.927225 1.081952 5.386134 3.225684 3.263078 4.291114 4.006264 3.880988 4.549628 5.625973 2.810439 4.602952 2.038267 2.535204 2.342214 3.681017 6.587096 6.732946 3.720131 2.402626 2.894096 2.871475 3.66951 2.905399 2.556755 3.16915 4.83849 5.970275 4.680592 3.041588 6.317036 4.75787 3.592208 3.688122 5.885241 7.149279 2.785289 3.85711 7.13676 3.93791 2.924254 4.307578 3.905605 2.522308 3.784355 2.595471 5.093715 6.844539 4.767503 3.119982	7.539525 3.498279 4.608205 2.969231 4.166654 6.824673 3.893933 2.82423 4.199148 4.288629 4.150698 1.518067 3.851391
High i=3	3.114865 2.639375 1.815265 3.97153 3.157677 6.888479 4.333933 4.131785 3.61824 4.224296 4.478117 6.83641	4.47687 1.673707 2.661594 4.031733 4.874429 4.566723 9.872442 2.367734 0.689261 3.831678 7.261367 2.865385 5.281288 5.810328 1.164427 4.582451 4.436615 5.760681	10.3551 4.517826 1.894344 -1.52287 2.247237 4.272311 2.167617 3.377569 1.389922 5.262377 3.019517 1.426379 2.337298 10.44505 3.424855 3.271532 4.045484 1.985204

2-Factor ANOVA method, tests the following:

- Are the mean treatments equal?
- Is there interaction between CO_2 change and temperature change?

2-Factor Anova Method:
H'o: No significant difference between Rate of CO_2 Change Levels [$\alpha_1 = \alpha_2 = \alpha_3$] H'a: Significant difference between Rate of CO_2 Change Levels [atleast one $\alpha_i \neq 0$]
H''o: No significant difference between Rate of Temperature Change Levels [$\beta_1 = \beta_2 = \beta_3 = 0$] H''a: Significant difference between Rate of Temperature Change Levels [atleast one $\beta_j \neq 0$]
H'''o: Temperature and CO_2 Change Levels don't interact [$(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{33} = 0$] H'''a: Temperature and CO_2 Change Levels interact [atleast one $(\alpha\beta)_{ij} \neq 0$]
Critical region: Reject Ho if P(f) < 0.05
Running Factorial Design Analysis in Minitab gives:

General Factorial Regression: Change in GDP versus Change in Temperature, Change in CO2 Emissions

Factor Information

Factor	Levels	Values
Change in Temperature	3	H, L, M
Change in CO2 Emissions	3	H, L, M

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	8	73.268	9.1585	2.68	0.008
Linear	4	56.173	14.0432	4.10	0.003
Change in Temperature	2	1.117	0.5586	0.16	0.850
Change in CO2 Emissions	2	54.869	27.4343	8.02	0.000
2-Way Interactions	4	8.001	2.0002	0.58	0.674
Change in Temperature*Change in CO2 Emissions	4	8.001	2.0002	0.58	0.674
Error	182	622.761	3.4218		
Total	190	696.028			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.84980	10.53%	6.59%	1.35%

Results:

- Reject H'o: Significant difference between Rate of CO_2 Change Levels
- Fail to Reject H''o: No significant difference between Rate of Temperature Change Levels
- Fail to Reject H'''o: No significant interaction between Temperature and CO_2 Change Levels

3.2 Discussion of Results: Effects and Interaction Plots

Figure 4 verifies the ANOVA the results that Change in CO_2 Levels has significant effect on the response variable

Figure 5 shows that

- highest Rate of GDP increase happens at High Rate of CO_2 Change
- there is low significant difference in the mean between High and Medium High Rates of CO_2 Change
- lowest Rate of GDP increase happens at low Rate of CO_2 Change
- GDP does start to slightly decline with increase in Change in Temperature from Medium to High

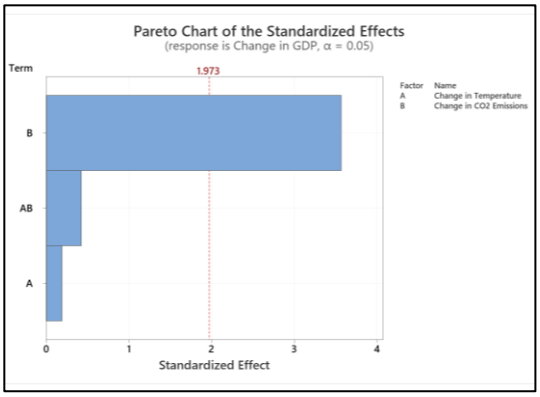

Figure 4: Effects Graph

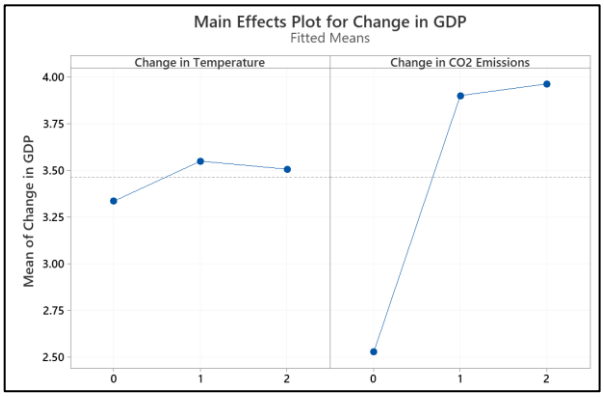

Figure 5: Main Effects Plot

Figure 6 shows Medium and Low Levels of Rate of CO_2 Change don't interact, as seen by their parallel lines. Between High and Medium Levels of Rate of CO_2 Change, there is interaction. For Medium Level,

GDP and Rate of Temperature Change have a slight linearly proportional relation, but for High Level, GDP starts to then go down with increase in Rate of Temperature Change.

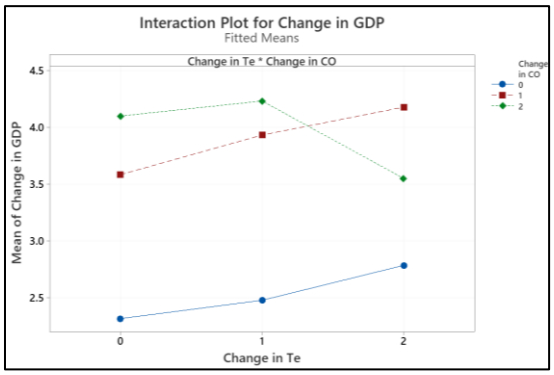

Figure 6: Interaction Plot

This can be summarized as:

- The best compromise between economic growth and CO_2 levels is keeping the per capita Rate of CO_2 increase at a mid-level.
- There is no significant benefit of increasing Rate of CO_2 Change from Medium to High levels.
- Infact, countries with High Level of Rate of CO_2 Change and High Level of Rate of Temperature Change are seeing decline in their GDP growth.
- Not all countries who are seeing higher levels of Rate of Temperature Change have citizens who are producing higher levels of CO_2. Many countries are seeing the negative consequences of climate change because of other countries producing too much CO_2.

4. DATA EVALUATION

To test for Type I errors, Duncan's Multiple Range-Test $R_p = r_p\sqrt{\frac{s^2}{n}}$. and Tukey's Procedure $q(\alpha, k, v)\sqrt{\frac{s^2}{n}}$. were done with Rate of CO_2 Change Levels as 3 treatments with 3 mean observations each.

Table 3: Type I Error Evaluation

p	2	3			
rp α=0.05 v=k(n-1) = 3(3-1) = 6	3.4610	3.5870			
Rp	3.6963	3.8309			
q α=0.05 v=k(n-1) = 3(3-1) = 6	4.3400				
percentile point	4.6351				
means of p samples, ascending	ybar1	ybar2	ybar3		
	2.3157	3.6286	4.1008		
				DUNCANS	TUKEYS
	difference of means	p	Rp	Significantly Different?	Significantly Different?
ybar1-ybar2	1.3129	2	3.6963	NO	NO
ybar2-ybar3	0.4722	2	3.6963	NO	NO
ybar1-ybar3	1.7851	3	3.8309	NO	NO

Although both tests show the difference between the CO_2 Levels are not significant, it is important to remember that the data is dealing with exceedingly small changes on a global scale, with many other factors.

- The $\mathbf{R^2}$ of GDP change (%) and CO_2 change (metric ton per capita) is 0.0361. Given the number of factors that effect a country's GDP, a 3.61% impact on the variation by just 1 metric ton per capita is very significant.
- The **correlation coefficient** of GDP change and Temperature change is only -0.005, but that means there is some negative effect. Studying change in Agricultural GDP only would give a clearer view.
- The $\mathbf{R^2}$ of Temperature change (°C) and CO_2 change (metric tons per capita) is 0.0004, which means an increase in 1 metric ton of CO_2 by 1 person accounts for 0.04% of the variance temperature increase, which is quite significant.

5. CONCLUSION

The 2-Factor Anova test proved that there is significant difference between the treatment effects caused by High, Medium and Low Rate of CO_2 Emission Change on Rate of GDP Change. The best compromise between economic development and CO_2 emissions is limiting the per capita emissions to a Medium level. There is no significant GDP improvement rate between Medium and High emission increase rate. Countries should put measures in place to limit emission rates based on how many citizens they have.

DATA SOURCES

- https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html
- https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html
- https://www.bea.gov/news/2017/gross-domestic-product-state-4th-quarter-2016-and-annual-2016
- zillow.com
- https://www.dol.gov/agencies/whd/mw-consolidated
- https://ncses.nsf.gov/indicators/states/indicator/bachelors-degree-holders-per-25-44-year-olds/table
- https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/topic-pages/tables/table-4
- https://climateknowledgeportal.worldbank.org/download-data
- https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?end=2016&start=1991&view=chart
- https://datatopics.worldbank.org/world-development-indicators/themes/environment.html

REFERENCES

- https://www.jchs.harvard.edu/sites/jchs.harvard.edu/files/05_harvard_jchs_americas_rental_housing_2017.pdf
- https://rstudio-pubs-static.s3.amazonaws.com/348670_ad694ac3e3cc475db019daa584ba663e.html
- https://www.un.org/en/sections/issues-depth/climate-change/
- https://www.bbc.com/news/world-africa-13317174