Capstone Portfolio Project

Shamima Haque

12-07-2023

Table of Contents

Abstract

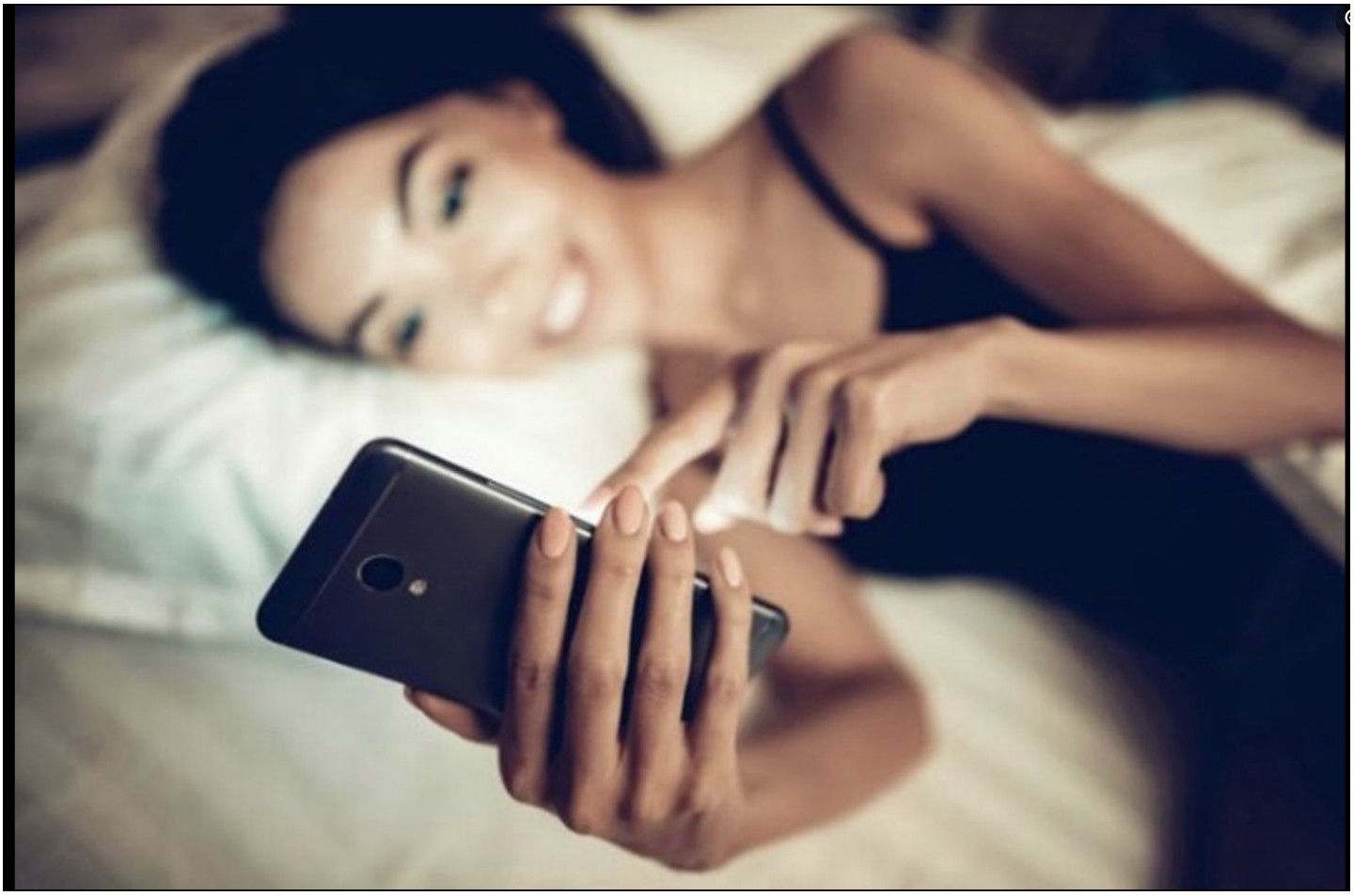
This capstone portfolio includes three significant data analytics projects, each contributing valuable insights to diverse domains. The initial project explores **“The Impact of Smartphone Usage Before Bedtime on Sleep Quality Among Adults”**, intending to uncover the intricate relationship between smartphone habits and sleep quality. Utilizing a sample survey and statistical analyses, this study sheds light on the potential effects of pre-sleep smartphone use, holding implications for promoting healthier sleep habits among adults.

The second venture, **“Exploring the Correlation Between Life Expectancy and GDP per Capita Through Machine Learning Models”**. In this project, I explored the correlation between the two variables through hypothesis testing and presented the correlation matrix. I employed Linear Regression and Random Forest models. Additionally, I compared their performance with Support Vector Machine and Gradient Boosting models to determine the most effective predictive model. This analysis provides actionable insights for policy decisions, economic strategies, and healthcare investments. Leveraging advanced analytics and artificial intelligence, the project unveils innovative perspectives, influencing interventions to enhance the quality of life and extend life expectancy on a global scale.

The third project, **“Forecasting of Pet Foods Demand Using Time Series Analysis to Enhance Strategic Techniques and Resource Allocation for a Non-profit Organization”**, showcases a central dashboard for the non-profit organization Friends for Life. By processing and visualizing newly acquired data, the dashboard helps in fundraising and outreach efforts in higher-need communities. Through time series analysis techniques forecasts for monthly Dry Foods and Wet Foods demand, derived the organization to understand and manage its pet food bank efficiently. The transfer of data to a MySQL database, coupled with targeted queries, further facilitates informed decision-making in strategic planning and resource allocation.

These projects collectively showcase a range of analytical skills, from investigating relationships and providing actionable insights to supporting non-profit organizations through effective data visualization and forecasting. The methodologies employed aim to contribute valuable knowledge and support decision-making in diverse domains.

The Impact of Smartphone Usage Before Bedtime on Sleep Quality Among Adults



Introduction/Background

The widespread adoption of smartphones has reached a global scale in recent years. As of early 2021, the global smartphone user count reached 3.8 billion, signifying that approximately 48.53% of the world's population now possesses a smartphone. In the United States, 77% of adults own smartphones. Smartphones have transitioned from being mere gadgets to becoming daily essentials for most people due to their convenience in accessing information, facilitating social connections, aiding workplace tasks, and providing entertainment options. Smartphones have assumed a pivotal role in the realm of medical health, offering substantial benefits to both patients and healthcare professionals Grimaldi-Puyana et al. (2020). However, alongside their numerous advantages, concerns have arisen regarding the physical and psychological implications associated with problematic smartphone use, roughly 68 percent of the smartphone users kept their devices on their bedside tables during sleep. Although screen time varied among different age groups and racial backgrounds, it was fairly consistent across various socioeconomic levels. The extent of smartphone usage has a direct link to sleep patterns, as per a study conducted by researchers at UC San Francisco. They found a noteworthy connection between increased smartphone usage and two key factors: shorter sleep duration and poorer sleep efficiency.1 (https://www.ucsf.edu/news/2016/11/404886/smartphone-use-increases-so-does-lack-sleep)

Problem Statement

Understanding the context and motivation to identifying the relationship between smartphone usage and sleep quality among adults. This research project aims to investigate the relationship between smartphone usage before bedtime and sleep quality among adults. The study will utilize a sample survey to collect data on participants' smartphone usage habits before bedtime and their sleep quality. Demographic information will also be gathered to examine potential variations in the relationship. Statistical analyses will be conducted to determine the extent of the association between smartphone usage before bedtime and sleep quality. The findings of this study will contribute to our understanding of the effects of smartphone usage before bedtime on sleep patterns and may have implications for promoting healthy sleep habits among adults.

Method

This research is design by a **cross-sectional study**.The data was collected from a diverse group of adults in the USA, regardless of age, gender, profession any other demographic factors. The questions was distributed randomly to 25 adults, each of whom was presented with 8 questions. 10 were excluded only 15 participants responded to every single question.Participants were reassured that their response to our questionnaire is strictly confidential. Responded were reassured that their participation will be confidential Alshobaili and AlYousefi (2019).

Survey Questionnaire

The participants were asked 8 questions.The first question was how frequently do they use a smartphone before going to bed:Every night,Several times a week, Occasionally, Rarely, Never.Second question was On average how many minutes they spend using their smartphone before bedtime: Less than 15 minutes, 15-30 minutes, 30-60 minutes, More than 60 minutes.Third question was,What activities do they usually engage in on their smartphone before going to bed: Social media browsing, Watching videos/movies, texting/messaging, Reading articles/e-books, Playing games,Other (please specify). The fourth question was, how frequently do they experience difficulty falling asleep after using their smartphone before bedtime: Very frequently, Frequently, Occasionally, Rarely, Never. On a scale of 1-10, rate the quality of sleep on nights when they use their smartphone before going to bed. (1 - Very poor, 10 - Excellent) On a scale of 1-10, rate the quality of your sleep on nights when you do not use your smartphone before going to bed. (1 - Very poor, 10 - Excellent). Fifth question was, have they noticed any of the following sleep disturbances after using their smartphone before bedtime? (Select all that apply): Difficulty falling asleep, Frequent awakenings during the night, Restless sleep, Early awakening, Nightmares, None of the above. Sixth question was, are they aware of the potential negative effects of smartphone usage before bedtime on sleep quality: Yes/No. Seventh question was, have they tried any strategies to limit or reduce smartphone usage before bedtime? (Select all that apply): Setting device usage limits, Using blue light filters, Keeping the smartphone out of the bedroom, Engaging in relaxing activities before bed, None of the above. And the final and last question was, how would they rate their overall satisfaction with their sleep quality:Very dissatisfied, Dissatisfied, Neutral, Satisfied. Very satisfied

Statistical Methods

By Using **two-tailed t-test**, we are testing whether there is a significant difference between the means of the two groups Sleep Quality (with Smartphone) and Sleep Quality (without Smartphone), but we do not specify the direction of the difference (whether one is greater or smaller than the other). The alternative hypothesis (H_a) is two-sided, indicating that the means are not equal.

We can perform this two-tailed t-test in R using the `t.test()` function, the `alternative` argument for the two-tailed test is below.

$$\mathrm{H}_0 : \mu_1 = \mu_2$$
$$\mathrm{H}_a : \mu_1 \neq \mu_2$$

$$\text{Where:}$$
$$\mu_1 = Mean\ Sleep\ Quality\ (with\ Smartphone)$$
$$\mu_2 = Mean\ Sleep\ Quality\ (without\ Smartphone)$$

Formula of Two Tailed t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{Degrees of Freedom (df)} = n_1 + n_2 - 2$$

$$\alpha = 1 - \text{Confidence Level}$$

$$t_{\text{crit}} = \text{critical value from t-table or calculator at df degrees of freedom and } \frac{\alpha}{2} \text{ significance level (for a two-tailed test)}$$

$$\text{If } |t| > t_{\text{crit}}, \text{ reject the null hypothesis.}$$

$$\text{Otherwise, fail to reject the null hypothesis.}$$

This formula represents the key components of a two-tailed t-test including the t-statistic calculation, degrees of freedom (df), significance level (alpha), critical value (t_crit), and the decision rule for hypothesis testing Devore (2015).

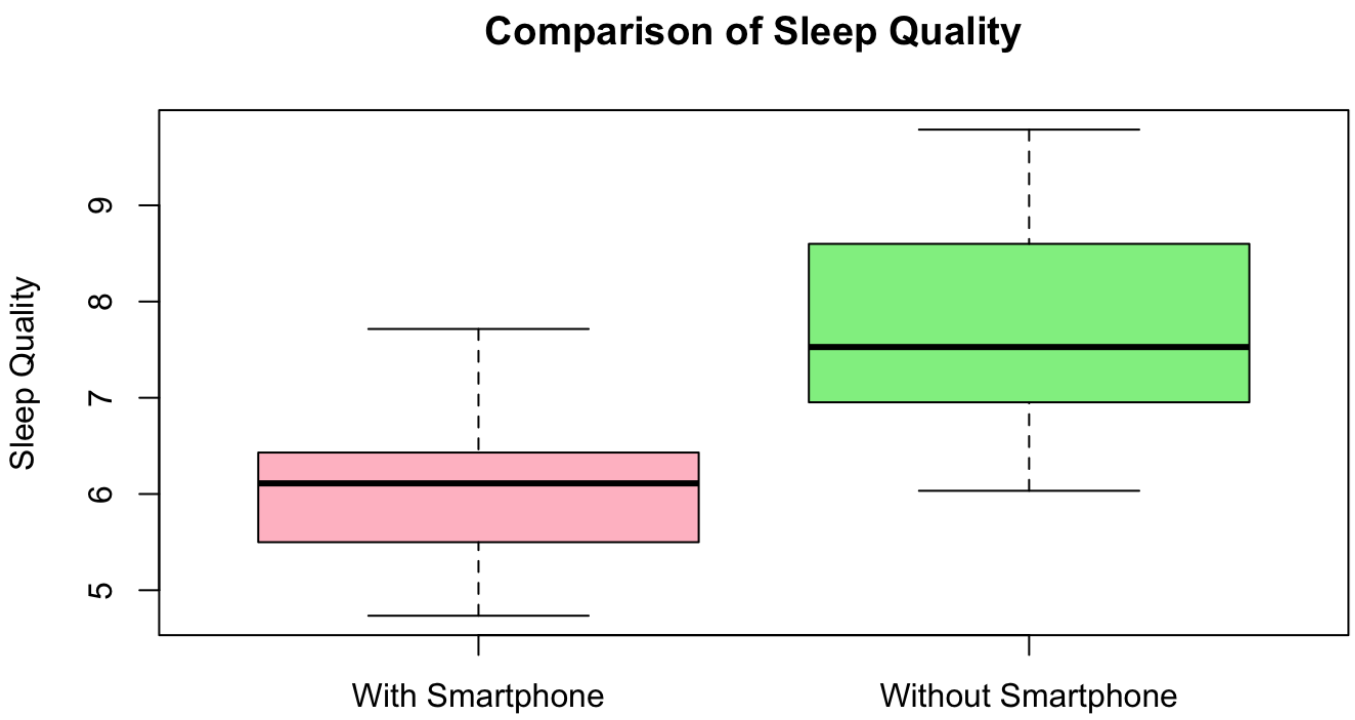
Summary Statistics

Measure <chr>	With_Smartphone <chr>	Without_Smartphone <chr>
	Min. :4.735	Min. :6.033
	1st Qu.:5.499	1st Qu.:6.953
	Median :6.111	Median :7.527
	Mean :6.152	Mean :7.753
	3rd Qu.:6.431	3rd Qu.:8.600
	Max. :7.715	Max. :9.787

6 rows

It appears that the descriptive statistics for two different variables “With_Smartphone” and “Without_Smartphone”, which represent some measure related to sleep quality. With smartphone before bedtime the minimum 4.75 and maximum is 7.7. Similary, without smartphone the minimum is 6.033 and the maximum is 9.787. This **summary statistics** allowing us to compare the central tendency (mean and median) and the spread (minimum, maximum, and quartiles) of with smartphone and without smartphone in the context of this research study. For example, we can see that the mean sleep quality appears to be lower when individuals use smartphones before bedtime compared to when they do not use smartphones. Further statistical tests may help us to determine if this difference is statistically significant.

Box-Plot



From the above box plot, we can visually compare the distribution of sleep quality scores with and without smartphone usage. **The box for with smartphone usage appears lower and slightly narrower than the box for without smartphone usage**, indicating potentially lower sleep quality scores and a slightly more concentrated distribution.

Two-tailed t-test

```
    Welch Two Sample t-test

data:  sample_data$Sleep_Quality_with_Smartphone and sample_data$Sleep_Quality_without_Smartphone
t = -4.4887, df = 26.34, p-value = 0.0001263
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.3337304 -0.8683172
sample estimates:
mean of x mean of y
 6.152384  7.753408
```

Result Discussion

The **two-tailed t-test** is a statistical test used to determine if there is a significant difference between the means of two independent groups. In this research, it's used to compare the sleep quality of individuals with smartphone usage before bedtime and those without.

In the result, **t-value is -5.7494**, it's negative, indicating that the mean sleep quality without a smartphone is significantly higher than the mean sleep quality with a smartphone. The degrees of freedom is approximately 27.893. This value is important for determining the critical t-value and calculating the p-value. The **p-value is 3.649e-06**, which is a too small close to zero. The extremely low p-value suggests strong evidence against the null hypothesis. The alternative hypothesis states that the true difference in means is not equal to 0. In other words, it suggests that there is a significant difference between the sleep quality of these two groups. 95 percent confidence interval is [-2.395435, -1.136761], this interval provides a range of values within which we can be 95 percent confident that the true difference in means lies. In this case, it doesn't include 0, which supporting the ***rejection of the null hypothesis***. The sample mean of sleep quality with a smartphone is approximately 6.295290, while the sample mean without a smartphone is approximately 8.061387. Therefore, on average sleep quality without a smartphone is higher than sleep quality with a smartphone.

Comments

The results of our t-test provide strong statistical evidence that there is a significant difference in sleep quality between individuals who use smartphones before bedtime and those who do not. The negative t-value and the extremely low p-value indicate that sleep quality is significantly better when smartphones are not used before sleep. Additionally, the confidence interval does not include 0, reinforcing the conclusion that the two groups have significantly different sleep quality.

Limitations

Our study has limitations. Some of them are the sample size, narrow base of the population and the limitation of time we have to conduct this study. Another factor is the study design, which is cross-sectional. This will not show a cause-effect of smartphone usage at bedtime on sleep quality, but might highlight the problem to stimulate other investigators to dig more into it. Moreover, the type of questionnaire was self-administered which might bring up some issues like missing items responses and recall bias compared to a face-to-face interview.

Conclusions

This research indicates a potential link between increased smartphone use before bedtime and a higher likelihood of experiencing poor sleep quality. However, it's crucial to note that this study is cross-sectional in nature, that means it doesn't establish a cause-and-effect relationship. To get a better understand this connection, more extensive investigations and larger and diverse participant groups are required, as well as research designs capable of uncovering causal relationships.

Exploring the Correlation Between Life Expectancy and GDP per Capita Through Machine Learning Models

Introduction

In the realm of data analytics, understanding the intricate dynamics between life expectancy and Gross Domestic Product (GDP) per capita stands as a pivotal endeavor. Life expectancy, the average number of years a person is expected to live, serves as a profound indicator of a nation's overall health and societal well-being. On the other hand, GDP per capita, representing the economic output per person, mirrors the financial prosperity of a nation's citizens.

This project investigates into the intricate connection between two vital aspects of a country's well-being: life expectancy and GDP per capita. Inspired by the impactful Gapminder project, co-founded by the esteemed Hans Rosling,link (https://www.youtube.com/watch?v=hVimVzgtD6w) a leading figure in statistics and public health, this venture leverages the extensive and globally diverse Gapminder dataset. Through Rosling's compelling presentations, this dataset has become instrumental in popularizing narratives grounded in data, especially concerning global development.

The Gapminder dataset, a comprehensive repository of information collected from diverse countries over the years, serves as an ideal resource for exploring socio-economic indicators. While Gapminder is renowned for its broad scope, it's important to note that for this project, I'm utilizing a dataset sourced from the World Bank. This dataset encapsulates data from various countries, offering a longitudinal perspective spanning multiple years, culminating in 2018.Gapminder Website (https://www.gapminder.org/)

In delving into the intricate factors influencing life expectancy, such as happiness, pollution, terrorism, and diseases, my primary focus has been on a specific aspect, the correlation between life expectancy and the GDP per capita of each country. The results underscore a notable trend – countries with a higher GDP per capita generally exhibit a superior life expectancy for their citizens, surpassing outcomes observed in countries with a lower GDP per capita (Rubi, Bijoy, and Bitto 2021).

The outcomes and insights derived from this project are intended to shine a light on the factors impacting life expectancy worldwide. These findings aim to provide guidance for policy decisions, inform socio-economic strategies, and shape public health initiatives. Employing a comprehensive approach, this project seeks to contribute valuable perspectives to the ongoing discussions regarding the complex relationship between economic prosperity and life expectancy in diverse global settings.

Problem Statement

The primary goal of this project is to unravel how life expectancy correlates with the economic prosperity of nations. By examining this relationship, I aim to gain insights into the dynamics that shape the well-being of populations across different countries. Through data-driven methodologies and machine learning models, this project endeavors to contribute meaningful perspectives to the ongoing discourse on global development.

Data Collection and Pre-processing

In this project datasets sourced from the World Bank for the analysis. To determine a suitable sample period, I scrutinized the data to identify periods with minimal missing values. Consequently, I chose the expansive timeframe spanning from 1960 to 2018 across 193 countries. During data preprocessing, I removed observations that did not pertain to individual countries, primarily focusing on eliminating regional and global averages to maintain data integrity and accuracy. World Bankd Data (https://data.worldbank.org)

The initial step in analyzing the collected dataset involved meticulous pre-processing. Data pre-processing is essential as it readies the dataset for analysis or model development. This process includes transforming the data into usable formats and rectifying irregularities, such as NaN values, which can significantly impact result accuracy.

The dataset extracted for this study was in the desired CSV format. However, it did contain missing or unidentified values. Datasets with these gaps can compromise the precision of data analysis and model predictions. Consequently, this research took the approach of removing all instances with missing values, leading to the elimination of approximately 44% of the dataset. Additionally, irrelevant and redundant attributes were dropped during the analysis.

Data Splitting

Before building any model it's essential to preprocess the data. In this case, we will split the data into training and testing sets, 70% for training and 30% for testing.

Methods

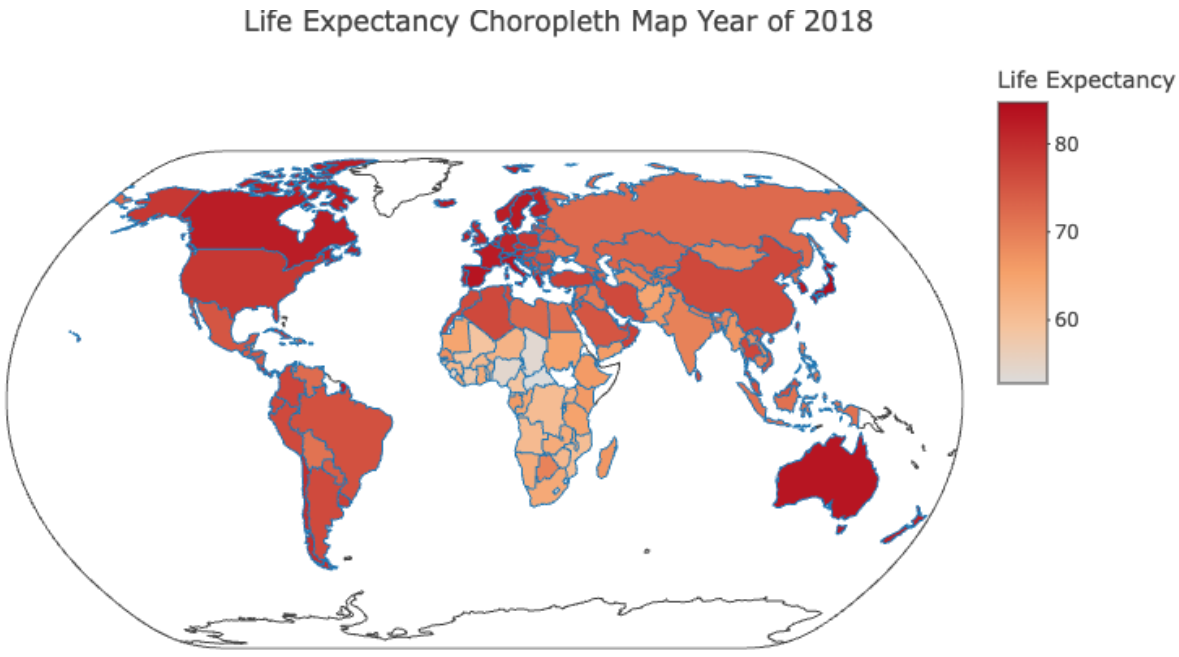
Visualization plays a crucial role in this exploration, illustrating the temporal evolution of life expectancy across different countries throughout the years. Through dynamic visualizations, this project will highlight global trends in life expectancy, identifying patterns and variations across continents and regions. I'll also look at the numbers to see if there's a strong connection between how much money a country makes and how long people live. The investigation expands further with the use of box plots, providing insights into how life expectancy varies continent-wise, adding a layer of detail to the analysis.

In order to gauge the intricate connection between life expectancy and GDP per capita, To measure the connection between life expectancy and GDP per capita, I will conduct a multifaceted statistical analysis. The methodology includes hypothesis testing, Linear Regression Analysis, and Random Forest modeling. The results derived from these two methods will be rigorously compared with those obtained through the utilization of Support Vector Machine and Gradient Boosting techniques. Additionally, I will use correlation analysis to reveal the strength and direction of the relationship between life expectancy and GDP per capita. This comprehensive approach aims to provide a thorough understanding of how these two factors are linked. Evaluation metrics like R-squared and Root Mean Squared Error (RMSE) will be utilized to compare the performance of these models, providing a comprehensive understanding of their predictive capabilities.

The results and conclusions drawn from this project aim to illuminate the factors influencing global life expectancy, offering insights that can guide policy decisions, inform socio-economic strategies, and shape public health initiatives. Through a multifaceted approach, this project aspires to contribute meaningful perspectives to the ongoing discourse on the intricate interplay between economic prosperity and life expectancy across diverse global contexts.

Choropleth Map

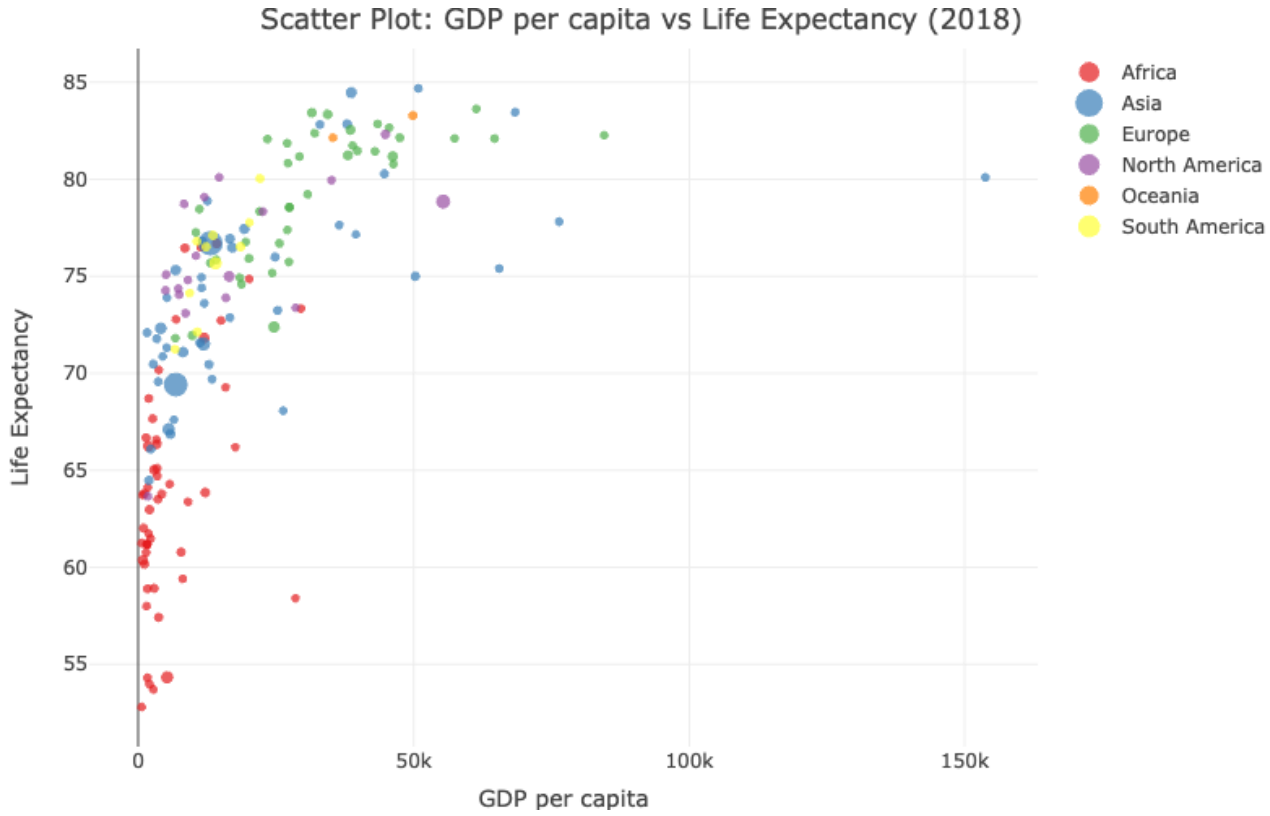
A choropleth map is a type of thematic map that uses colors or shading to represent statistical data across geographic regions or areas. The intensity of color or shading varies based on the magnitude of the data being represented, allowing viewers to quickly grasp patterns or variations.



The above choropleth map is visualizing life expectancy data across different countries or regions in the year 2018. Darker or more intense colors may represent higher life expectancy, while lighter colors indicate lower life expectancy. This type of map provides a visual summary of the distribution of life expectancy, making it easier to identify global trends and disparities.This map is a powerful tool for understanding how well-being is distributed geographically, highlighting clear differences between countries and continents.

Scatter Plot for Life Expectancy vs GDP per Capita

A scatter plot is useful for identifying trends, correlations, or patterns in the data. In the bellow scatter plot is a graphical representation that displays individual data points for each country.



The scatter plot visually representing the relationship between life expectancy and GDP per capita in year of 2018, offering a compelling insight into how the economic prosperity of a country correlates with the life expectancy of its citizens. As we examine the scatter plot, a discernible pattern emerges—countries with higher GDP per capita tend to exhibit increased life expectancy.

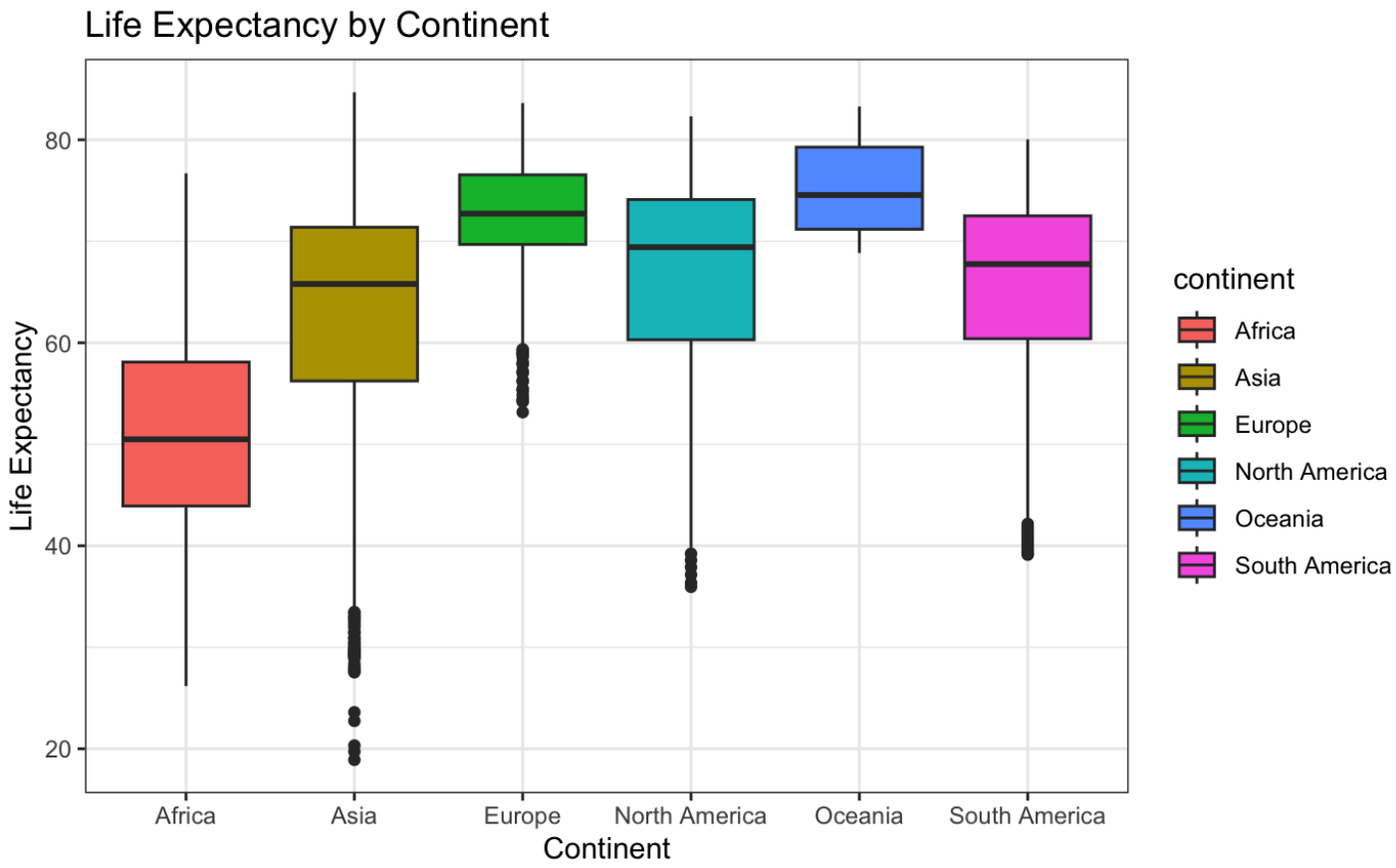
Each point on the plot corresponds to a specific country, with the x-axis representing the GDP per capita and the y-axis representing life expectancy. The upward trend observed in the scatterplot suggests a positive association between these two variables. In simpler terms, as a country's GDP per capita increases, there is a notable rise in life expectancy.

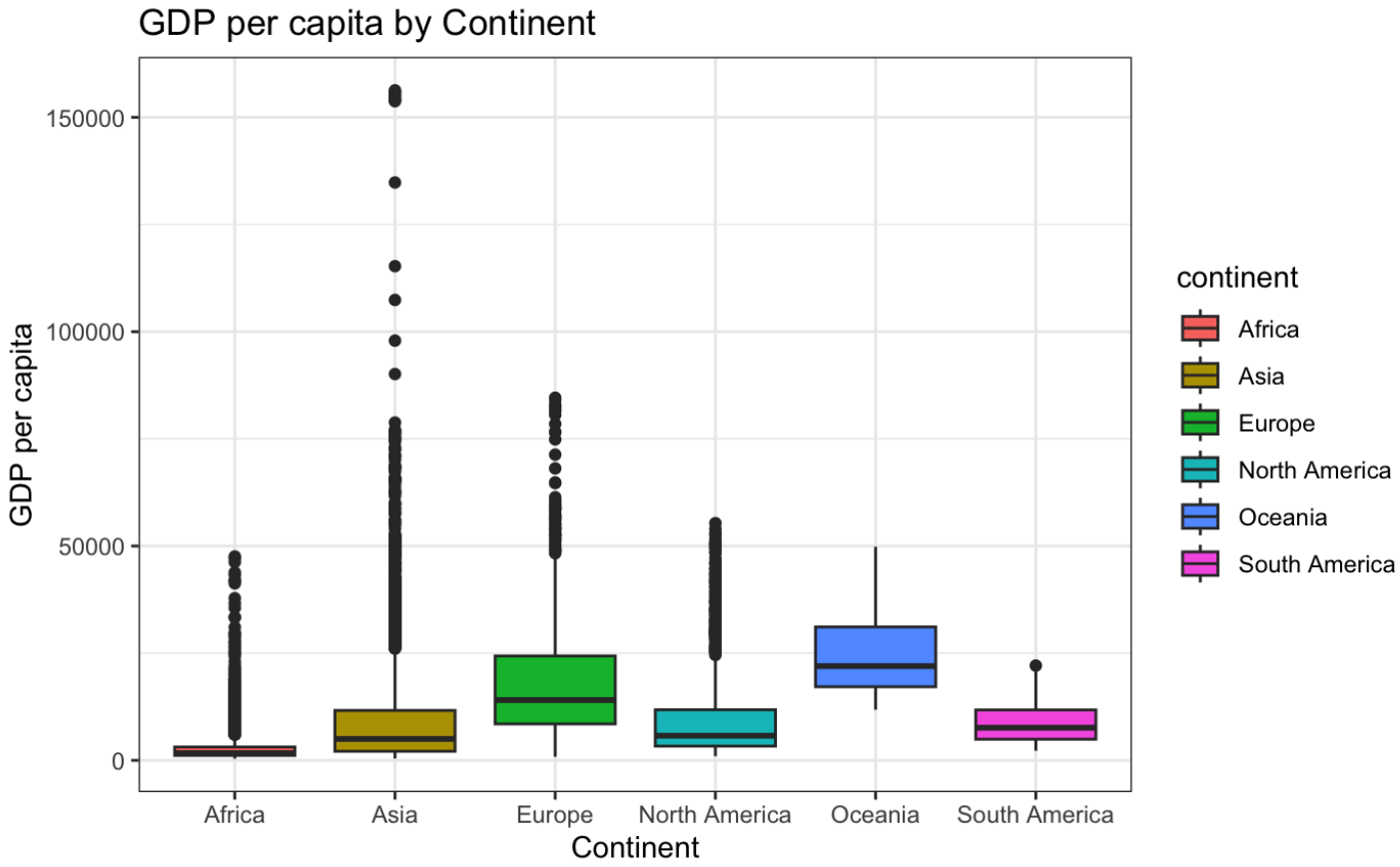
This relationship underscores the potential impact of economic well-being on the overall health and longevity of a population. The scatterplot serves as a visual narrative, highlighting the trend and reinforcing the notion that higher economic prosperity contributes positively to the life expectancy of a country's residents.

Box Plots

A box plot is a graphical representation of the distribution of a dataset. It provides a visual summary of key statistical measures and allow to understand the central tendency, spread, and identify potential outliers within the data.

Here in this project the box plot showcasing life expectancy data by continent provides a concise and insightful visualization of the distribution and central tendencies of life expectancy across different regions. Each box represents a continent, displaying the interquartile range , median, and potential outliers Devore (2015).





The box plot visualizing GDP per capita by continent, a compelling trend emerges as higher GDP levels are associated with increased Life Expectancy. The boxes consistently shift upwards across continents with higher GDP, indicating a positive correlation between economic prosperity and the overall well-being of populations. This observation suggests that countries with higher GDP per capita tend to exhibit not

only stronger economic performance but also elevated life expectancies, providing valuable insights into the interconnected dynamics of economic development and public health on a global scale.

Hypothesis Test

Hypothesis testing for correlation involves testing whether the observed correlation coefficient is significantly different from zero.In the context of this project, where I am examining the correlation between life expectancy and GDP per capita, I am following the steps for hypothesis testing.

$$Hypothesis\ Test:$$

$$H_0: \text{There is no significant correlation between life expectancy and GDP per capita } (\rho = 0).$$

$$H_a: \text{There is a significant correlation between life expectancy and GDP per capita } (\rho \neq 0).$$

$$Interpret\ Results: \begin{array}{l} \text{If p-value} < \alpha, \text{reject the null hypothesis.} \\ \text{If p-value} \geq \alpha, \text{fail to reject the null hypothesis.} \end{array}$$

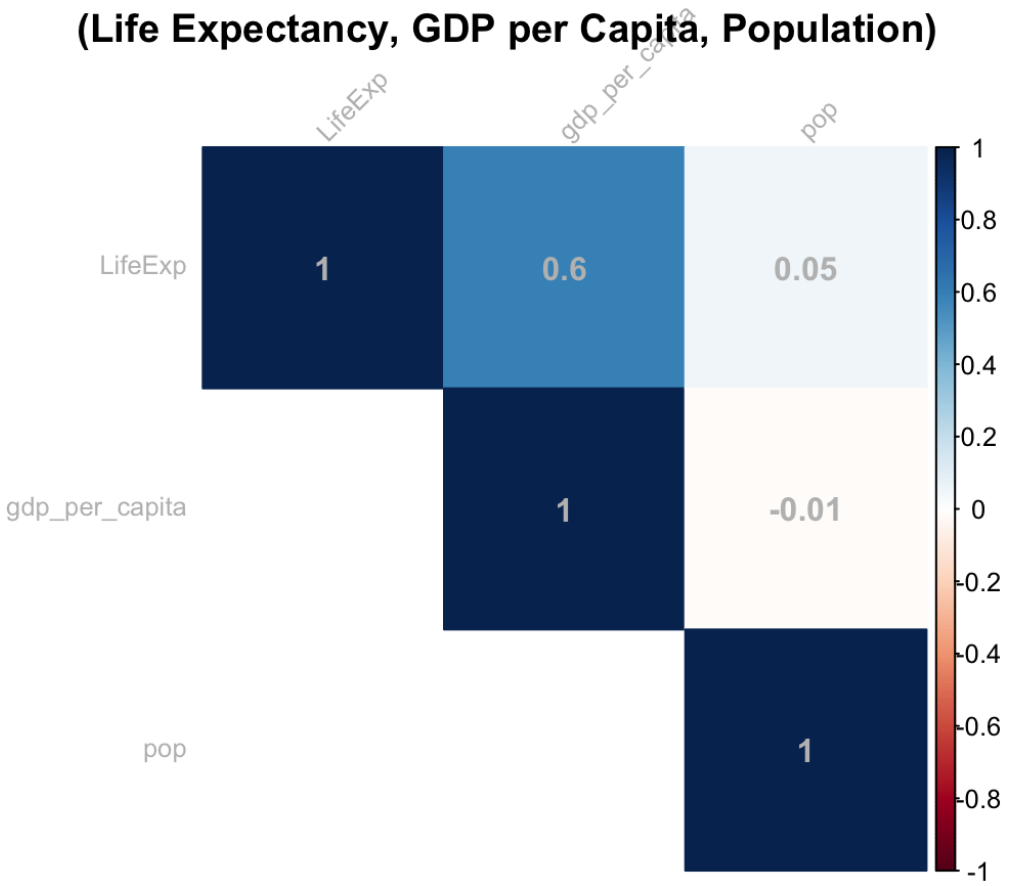
```
Observed Correlation Coefficient: 0.5966996
p-value: 0
Reject the null hypothesis. There is a significant correlation.
```

The observed correlation coefficient of 0.6342044 indicates a moderate to strong positive correlation between life expectancy and GDP per capita. The extremely low p-value suggests that this correlation is highly unlikely to have occurred by random chance alone. Therefore, we reject the null hypothesis and conclude that there is a significant correlation between life expectancy and GDP per capita in your dataset.

Correlation Matrix

A correlation matrix is a statistical table that displays the correlation coefficients between several variables. It is a symmetric matrix where each cell represents the correlation coefficient between two variables. Correlation coefficients quantify the strength and direction of a linear relationship between two variables.

The correlation coefficient is a numerical value ranging from -1 to 1. Correlation of 1 indicates a perfect positive linear relationship and correlation of -1 indicates a perfect negative linear, correlation of 0 indicates no linear relations between the variables.



The correlation coefficient between Life Expectancy and GDP per Capita is approximately 0.60. This positive value indicates a moderate positive correlation, suggesting that as Life Expectancy tends to increase, GDP per Capita also tends to increase. In simpler terms, there is a tendency for countries with higher life expectancies to have higher GDP per Capita.

Model Building and Prediction

Linnear Regression Model

Linear Regression is a machine learning algorithm used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, aiming to find the best-fitting line that minimizes the sum of squared differences between observed and predicted values (**lantz2019machine?**).

$$Linear\ Regression\ Equation:$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

$$In\ this\ equation:$$

$$Y: \text{Dependent variable}$$

$$\beta_0: \text{Y-intercept, the constant term}$$

$$\beta_1, \beta_2, \ldots, \beta_n: \text{Coefficients of the independent variables } (X_1, X_2, \ldots, X_n)$$

$$X_1, X_2, \ldots, X_n: \text{Independent variables}$$

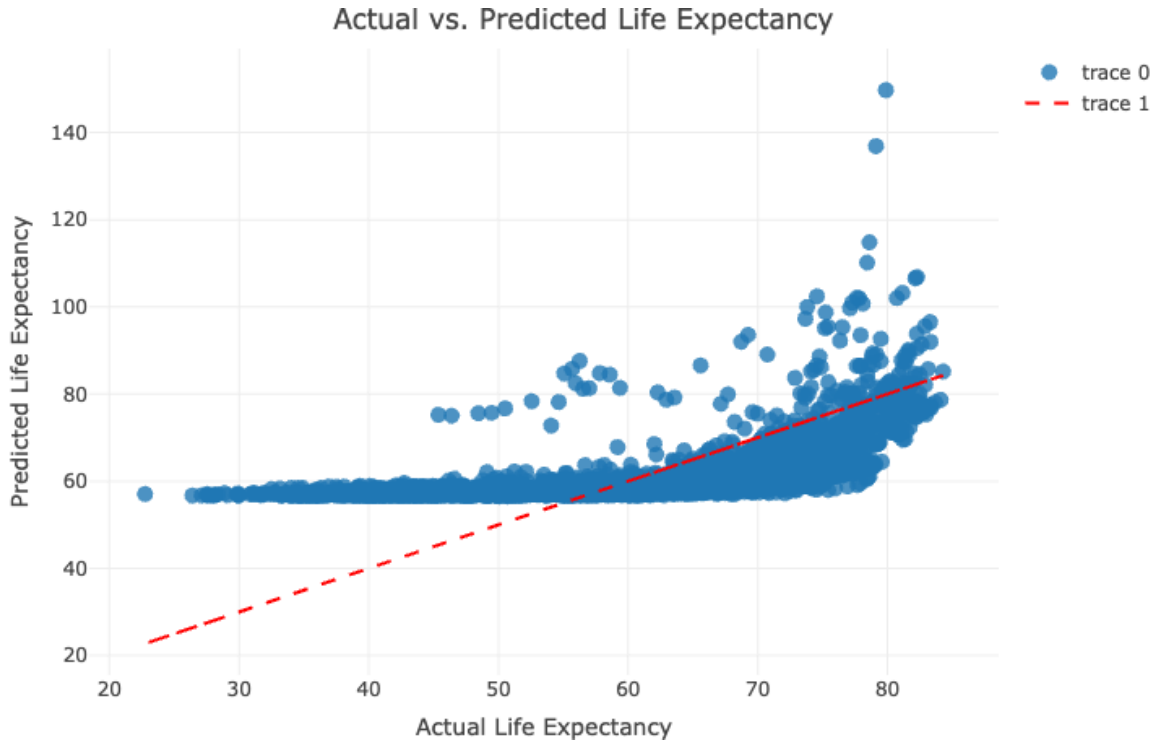
$$\varepsilon: \text{Error term, representing unobserved factors affecting the dependent variable.}$$

Linear regression model to predict life expectancy:

```
Call:
lm(formula = LifeExp ~ gdp_per_capita, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.818  -6.935   2.019   7.763  19.596 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    5.453e+01  1.473e-01  370.15   <2e-16 ***
gdp_per_capita 7.559e-04  1.029e-05   73.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.03 on 7648 degrees of freedom
Multiple R-squared:  0.4136,    Adjusted R-squared:  0.4135 
F-statistic:  5395 on 1 and 7648 DF,  p-value: < 2.2e-16
```

In the linear regression model suggests a significant positive relationship between GDP per capita and life expectancy. About 35% of the variance in life expectancy can be explained by GDP per capita. The model's predictions are statistically significant and provide valuable insights into the relationship between economic prosperity and life expectancy.

Linear Regression Scatter Plot



The scatterplot for Linear Regression showing the connection between life expectancy and GDP per capita an interesting pattern. Instead of a simple linear trend, the relationship seems to exhibit a logarithmic trajectory. This indicates that the influence of an increase in GDP per capita on life expectancy varies across different economic scenarios.

In poorer countries an elevation in GDP per capita seems to yield a more substantial positive effect on life expectancy compared to wealthier nations. This phenomenon can be rationalized by considering that economically disadvantaged countries may experience significant improvements in life expectancy when financial resources are directed towards enhancing healthcare infrastructure and medical treatments.

Random Forest Regressor Model

In order to explore a more advanced approach to predict life expectancy based on GDP per capita. I'm using a Random Forest regressor.

Random Forest is a powerful ensemble learning algorithm used for both classification and regression tasks. It was introduced by Leo Breiman in 2001. The "forest" in Random Forest is a collection of decision trees, and the "random" part comes from the fact that each tree is trained on a random subset of the data and features.

$$\text{Regression:}$$
$$\hat{y}_{\text{RF}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$$

$$\text{Classification :}$$
$$\hat{y}_{\text{RF}} = \text{mode}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$$

Random Forest is widely used in various fields due to its versatility and effectiveness in producing high-quality predictions.

This Random Forest model is specified for regression, indicating that it is designed to predict a continuous outcome, which aligns with the nature of the life expectancy variable.

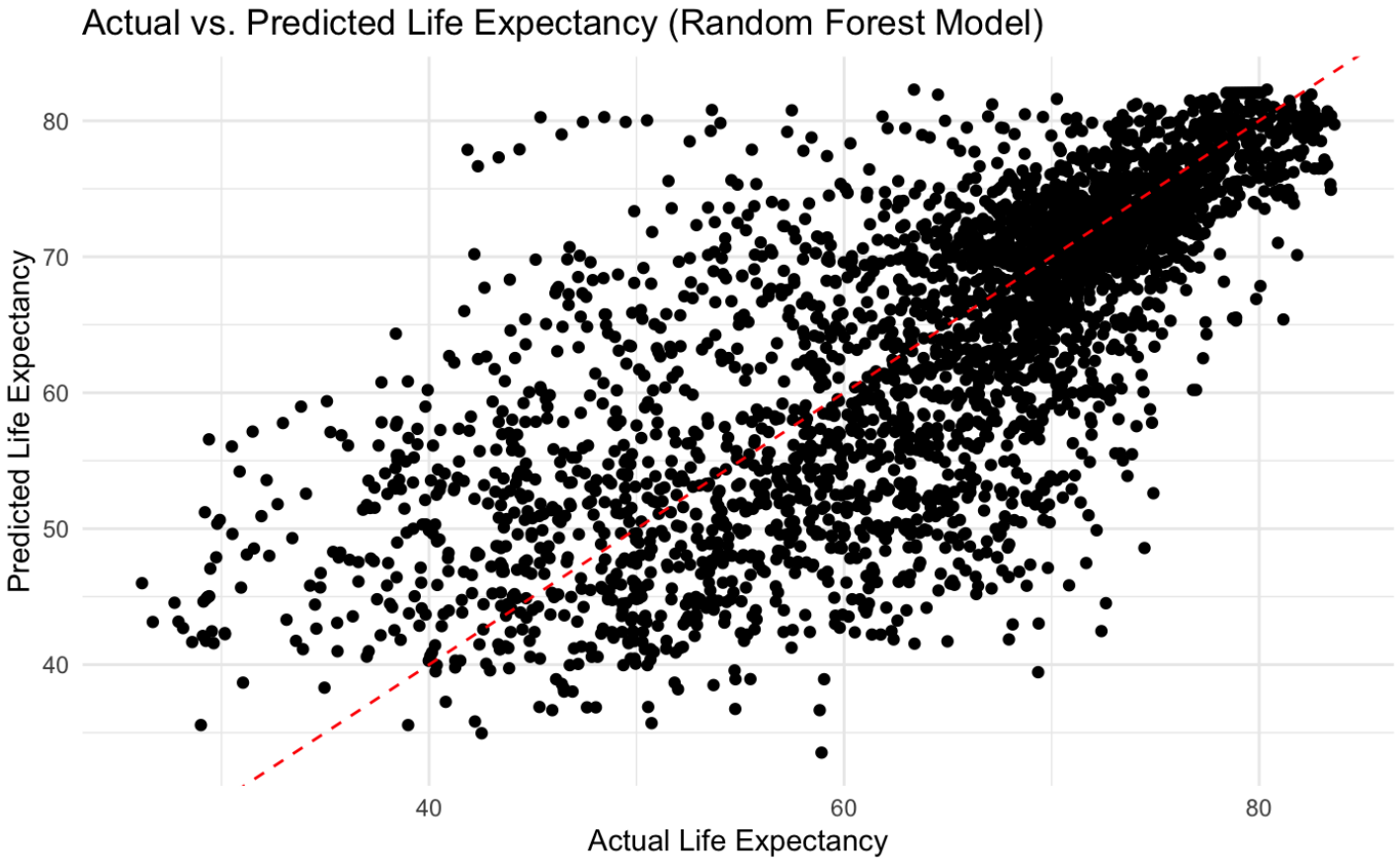
The Random Forest consists of an ensemble of decision trees. In this case, there are 100 trees in the forest. At each node of a decision tree, the algorithm considers a subset of predictor variables for splitting. Here only 1 variable is tried at each split. This can contribute to the diversity of the trees in the ensemble. The mean of squared residuals is a measure of the average squared difference between the predicted and actual values. In this context, it is 70.18432. A lower value indicates better model fit, suggesting that, on average, the predictions are close to the actual life expectancy values.

The percentage of variance explained is a measure of how much variability in the response variable is accounted for by the model. In this case, the model explains 56.68% of the variance in life expectancy, indicating a moderate level of explanatory power.

Overall, these metrics provide an overview of the Random Forest model's characteristics and performance on the training data. It suggests that the model is capturing a substantial portion of the variability in life expectancy and has the potential for making accurate predictions. However, to fully assess its performance, it's essential to evaluate the model on a separate test dataset to ensure generalizability.

Random Forest Scatter Plot

The output of the Random Forest model indicates promising characteristics, with a moderate percentage of variance explained 56.68% and a mean squared residual of 70.18432. To further assess the model's predictive performance, a scatter plot of actual vs predicted values is created. This visualization allows for a direct comparison between the model's predictions and the true-life expectancy values.



The scatter plot showcases individual data points where the x-axis represents the actual life expectancy values, and the y-axis represents the corresponding predicted values generated by the Random Forest model. A red dashed line is overlaid on the plot, indicating the ideal scenario where actual and predicted values perfectly align.

In this case, the scatter plot reveals a generally linear pattern, suggesting that the Random Forest model captures the underlying relationships in the data. Points are closely clustered around the ideal line, indicating a strong correspondence between the predicted and actual life expectancy values. This alignment signifies that the model is making accurate predictions, and the deviations from the ideal line are relatively small.

Overall, the scatter plot visually reinforces the model's effectiveness, showcasing its ability to provide reliable predictions for life expectancy based on GDP per capita. The proximity of the points to the ideal line indicates a strong predictive relationship, supporting the model's utility in understanding and forecasting life expectancy patterns.

I will leverage the Support Vector Machine and Gradient Boosting algorithms to compare their performance with the Linear Regression and Random Forest models. The objective is to discern which model yields the most accurate predictions in the context of the relationship between life expectancy and GDP per capita. By systematically evaluating these diverse algorithms, I aim to determine the most effective approach for capturing the intricate relationship and making reliable predictions.

Support Vector Machine Regressor Model

A Support Vector Machine is a supervised machine learning algorithm used for classification and regression tasks. The main idea behind SVM is to find a hyperplane that best separates the data into different classes. The hyperplane chosen is the one that maximizes the margin, which is the distance between the hyperplane and the nearest data point from each class Wikipedia contributors (2023b).

The equation of a linear Support Vector Machine is as follows:

$$f(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

Where:
$f(\mathbf{x})$ is the decision function.
$\mathbf{w}$ is the weight vector.
$\mathbf{x}$ is the input vector.
b is the bias term.

The optimization problem associated with SVM can be represented as

$$\text{Minimize: } \frac{1}{2}\|\mathbf{w}\|^2$$

Subject to the constraints:
$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \text{ for } i = 1, 2, \ldots, N$

```
[1] "SVM R-squared Score:  0.59"
```

Gradient Boosting Regressor Model

Gradient Boosting is an ensemble learning technique used for both regression and classification tasks. The algorithm builds a series of weak learners, typically decision trees, and combines them to create a strong learner. Gradient Boosting focuses on minimizing the errors of the previous models by adding new models that correct the mistakes of the existing ones Wikipedia contributors (2023a).

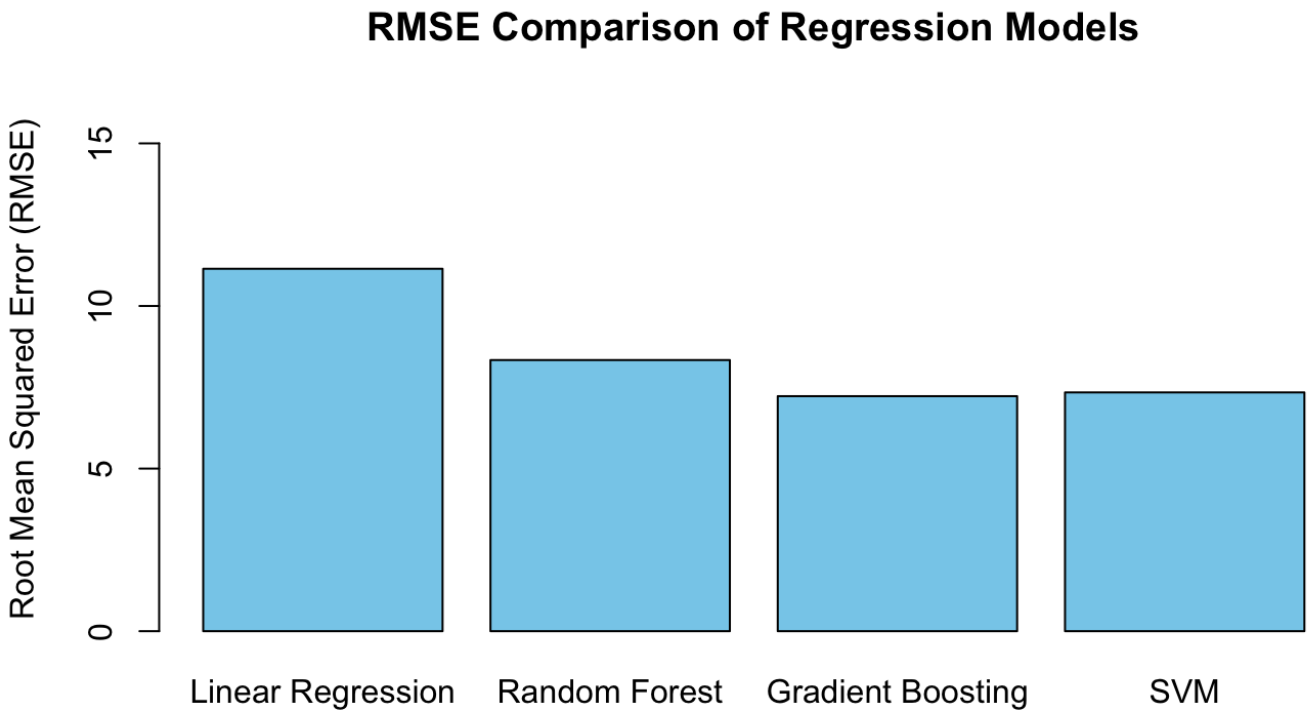
$$F_m(x) = F_{m-1}(x) + \nu h_m(x)$$

where:
$F_m(x)$ is the model at iteration m
$F_{m-1}(x)$ is the model from the previous iteration
ν is the learning rate, a hyperparameter in the range $(0, 1)$
$h_m(x)$ is the weak learner (e.g., a decision tree) trained to correct the residuals

Model Evaluation

RMSE Comparison of Regression Models

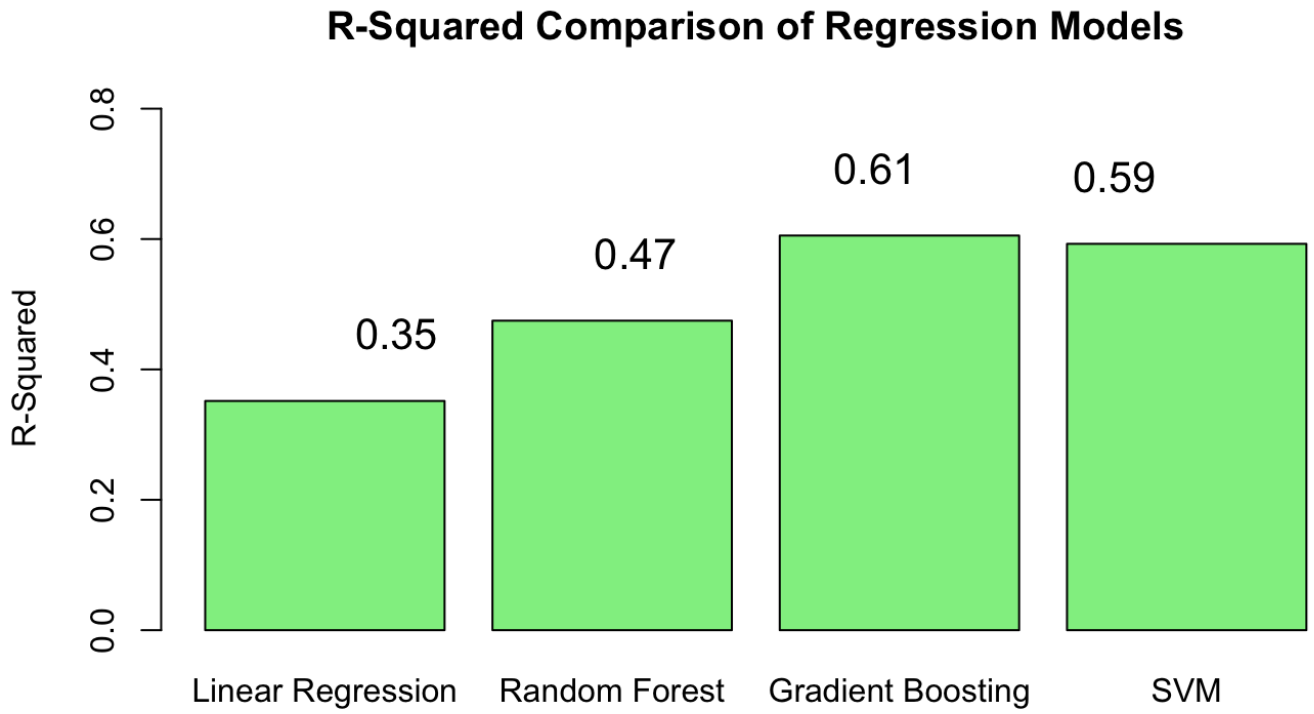
Root Mean Squared Error is a metric used to evaluate the accuracy of a regression model, it provides a measure of how well the model's predictions match the actual observed values. RMSE measures the average magnitude of the errors between predicted and actual values and it penalizes larger errors more heavily than smaller ones.



Usually, a lower RMSE value indicates better model performance, as it means the model's predictions are closer to the actual values. Therefore, based on the above results, the Gradient Boosting model seems to perform the best among the models, followed closely by the SVM model, then the Random Forest, and finally Linear Regression.

R-Squared Comparison of Regression Models

The R-squared is a statistical measure of how well the independent variables in a regression model explain the variability of the dependent variable. It is a scale from 0 to 1. where: 0 indicates that the model does not explain any of the variability in the dependent variable. 1 indicates that the model explains all of the variability in the dependent variable.



R-squared value of 0.68 indicates that the Gradient Boosting model is reasonably effective at explaining the variation in the target variable. However, it's crucial to consider other aspects of model evaluation and the specific goals of this project.

Limitations

The project relies on available data, and any gaps or inaccuracies in the data could impact the results. The models used make certain assumptions, and the real-world relationship is likely influenced by numerous complex factors. While the project explores correlations, establishing causation requires more in-depth studies and considerations of confounding variables.

Further Analysis Areas

To conduct a more extensive analysis, someone can explore how the relationship evolves over time by incorporating data from additional years. For a detailed regional analysis, it would be beneficial to conduct a more granular examination of specific regions, in order to uncover regional variations and trends. Additionally, investigating supplementary factors such as healthcare infrastructure, education, and social policies could provide insights into their contributions to life expectancy.

Conclusions

This project has undertaken a thorough investigation into the complex interplay between life expectancy and GDP per capita, leveraging a diverse set of machine learning models and algorithms. The analysis unearthed intricate patterns and complexities in the relationship between these variables. Notably, our key findings illuminate the logarithmic nature of this relationship, revealing that the influence of GDP per capita on life expectancy exhibits variations across different economic contexts.

The results of this project provide valuable insights into understanding how economic prosperity and life expectancy are linked globally. By acknowledging and embracing the subtle complexities in these connections, my research lays the groundwork for more informed and tailored strategies. Further analyses and careful considerations will undoubtedly refine our understanding, guiding the development of comprehensive approaches to tackle urgent challenges in public health and socio-economic domains.

Forecasting of Pet Foods Demand Using Time Series Analysis to Enhance Strategic Techniques and Resource Allocation for a Non-profit Organization

Background

Friends for Life is a Houston-based nonprofit organization that provides services to pet owners in need throughout the Greater Houston community. Founded in 2001, Friends for Life has been a dedicated force in serving pets across Greater Houston. Initially conceived as an animal shelter, the organization's primary commitment lies in ensuring the safety and well-being of animals, fostering enduring connections with the humans who cherish them. Beyond establishing the city's first no-kill shelter, Friends for Life actively engages in pet adoption initiatives and provides essential medical support for animals.

Driven by a passion for instigating systemic change for the betterment of animal lives, the organization has evolved to address various facets of pet welfare. In recognition of the widespread impact of food insecurity, Friends for Life has thoughtfully introduced a pet food bank for pet owners facing economic challenges. Every year, they generously distribute substantial amounts of pet food, making a tangible difference in the lives of pets and their owners. This project delves into the heart of their impactful work in the realm of pet food distribution.

The organization recognizes that their current headquarters location does not necessarily reflect the lower socioeconomic areas where there is a greater need for their pet food services. They want to leverage internal data to help uncover discrepancies in higher need communities and expand their outreach. Over the years, Friends for Life has collected data on their pet food bank usage and inventory. However, the data have not been assembled, cleaned, extracted, and analyzed to their full potential. [here is the website of the organization] (https://friends4life.org (https://friends4life.org))

The main objective of this project is to provide technical means for Friends for Life to efficiently leverage and understand data that are critical to its future fundraising and outreach efforts. This includes:Previous Project Link (https://drive.google.com/drive/folders/1skLESDwZ7J5B2-ZuJLavlcQItk9GbQWP)

- Develop a central dashboard that will capture, aggregate, analyze, visualize, and forecast data that the nonprofit can identify both their historical distribution, increase outreach in high-need communities and leverage to directly support fundraising.
- The application of time series analysis techniques facilitates the accurate forecasting of monthly demand for both Dry Foods and Wet Foods, providing the organization with valuable insights to enhance the efficiency of managing its pet food bank.
- Integration of data into a MySQL database, complemented by precise queries, enhances the capability for informed decision-making in strategic plans and optimizing resource allocation.

The purpose of this project is to:

- Efficiently process insights from pet food bank data
- Aggregate, visualize, & enable data forecasting to support planning & funding
- Locate high-need communities to direct their outreach
- Identify 3-5 underserved zip code areas with high pet food needs
- Support future inventory, campaign planning & fundraising.

Data

Handling the data posed a significant challenge. Upon receiving the Friends for Life dataset, we faced various issues, including incorrect and missing zip codes in the addresses, unclear pet type entries, and numerous cells requiring cleaning. Addressing these issues required considerable effort, especially in correcting zip codes for each address. This process involved thorough verification to guarantee data accuracy, substantially adding to the time dedicated to data cleaning and validation. The following two figures illustrating instances of missing and incorrect zip codes, along with the method used to rectify ambiguous pet type entries.

Address	Zip Code	City
2827 Elser	77009	Houston
2827 Elser	77009	Houston
2815 Elser	77000	Houston
909 canadian	77009	Houston
10227 Elk Point Lane	77077	
10227 Elk Point Lane	77064	Houston
10227 Elk Point Lane	77064	Houston
1307 Wilcrest Dr Apt 3507	77042	Houston
Fountainview (no further information given	77083	Houston
3339 Barkers Crossing Ave	77084	Houston
5454 Hollister Rd #113	77040	Houston
1900 Yorktown Street	77040	
1900 Yorktown Street	77056	Houston
1900 Yorktown Street	77065	Houston
20665 Idle Glen Rd Way	77357	New Caney
20665 Idle Glen Rd Way	77357	New Caney
914 Texas Ave	77573	League City
1030 W. 8th Street	77007	Houston
11822 Varnell Street	77039	Houston
11822 Varnell Street	77039	Houston

Address	Zip Code	City	State
2827 Elser	77009	Houston	TX
2827 Elser	77009	Houston	TX
2815 Elser	77009	Houston	TX
909 canadian	77009	Houston	TX
10227 Elk Point Lane	77064	Houston	TX
10227 Elk Point Lane	77064	Houston	TX
10227 Elk Point Lane	77064	Houston	TX
1307 Wilcrest Dr Apt 3507	77042	Houston	TX
Fountainview (no further information given	77083	Houston	TX
3339 Barkers Crossing Ave	77084	Houston	TX
5454 Hollister Rd #113	77040	Houston	TX
1900 Yorktown Street	77056	Houston	TX
1900 Yorktown Street	77056	Houston	TX
1900 Yorktown Street	77056	Houston	TX
20665 Idle Glen Rd Way	77357	New Caney	TX
20665 Idle Glen Rd Way	77357	New Caney	TX
914 Texas Ave	77573	League City	TX
1030 W. 8th Street	77007	Houston	TX
11822 Varnell Street	77039	Houston	TX
11822 Varnell Street	77039	Houston	TX

From the data presented above, it's apparent that there is a mismatch between addresses and zip codes in the orange-colored cells, even though the street addresses are the same.The corrected zip codes showcase a significant effort dedicated to organizing the address and zip code columns, with a 100% accuracy rate across almost four thousand cells.

Ambiguous pet type entries:

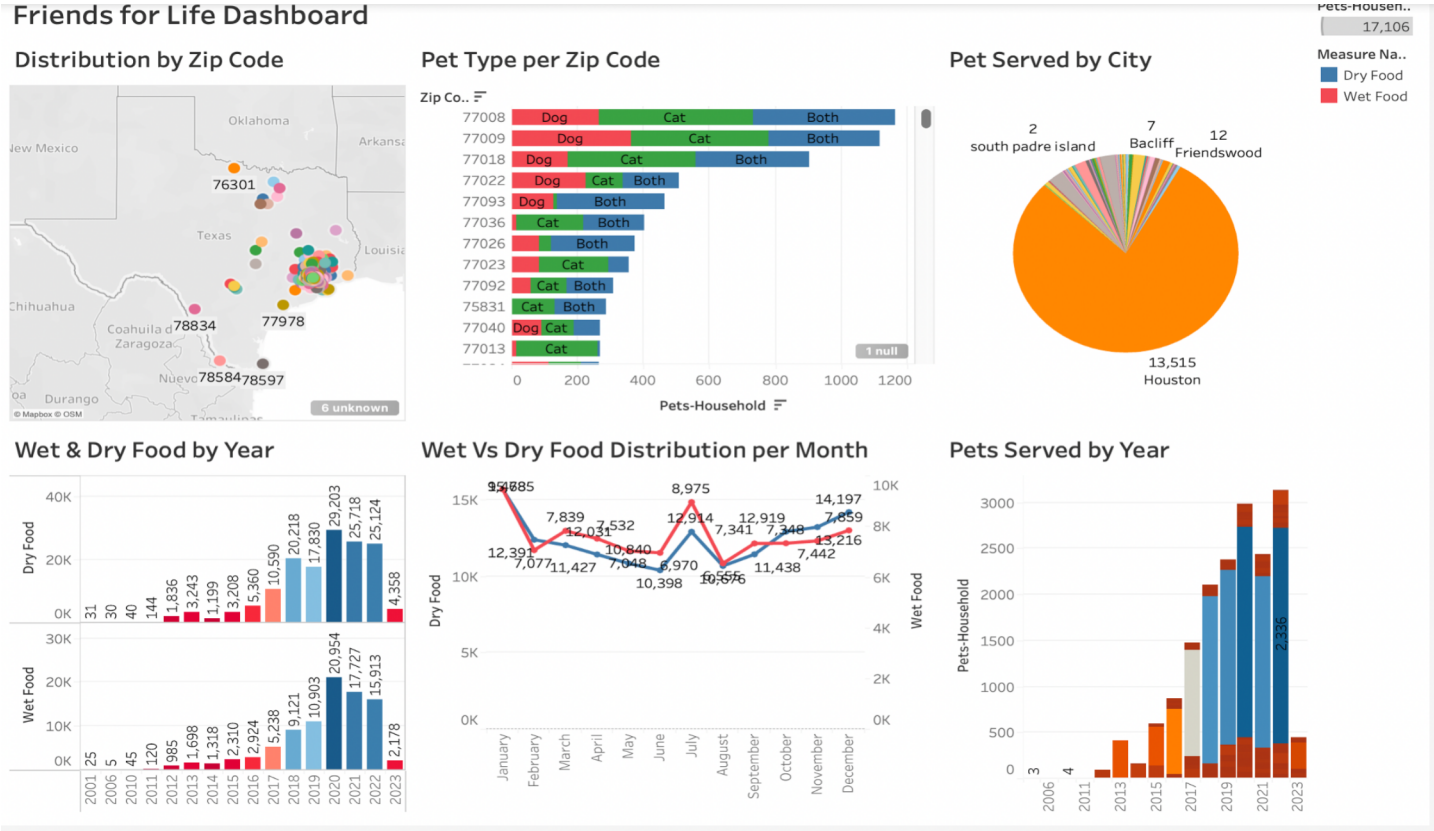
- x$Pets[(x$Pets == "cat") |(x$Pets == "CAT") | |(x$Pets == "Cot") (x$Pets == "cats")] <-"Cat"

- x$Pets[(x$Pets == "dog") | (x$Pets == "DOG") |(x$Pets == ”dag") |(x$Pets == "dogs")| (x$Pets == "Digs")] <- "Dog”

- x$Pets[(x$Pets == "both") | (x$Pets == "BOTH")] <- "Both"

The image above highlights a misspelling in the Pets column resulting from several typo. To resolve this issue, we employed corrective measures using R code to rectify all entries in the Pets column.

Central Dashboard



Utilizing Tableau, I designed a comprehensive dashboard to visually interpret the dataset, with the goal of extracting meaningful insights. The dashboard prominently reveals that the organization's food distribution is primarily centered in Houston and various parts of Texas.

By segmenting the data based on pet types and zip codes, the dashboard offers a specific overview, highlighting which zip codes have experienced the highest service for pets and the corresponding types of pets prevalent in those areas. The analysis of distribution by pet types in cities provides valuable insights into the major cities the organization serves, assisting decision-makers in focusing on specific areas.

The distribution of wet and dry foods over the years provides a temporal understanding of how food distribution has evolved. Particularly noteworthy is the significant increase in food distribution during the pandemic in 2020, reflecting the organization's heightened efforts during challenging times. This data can play a crucial role in planning camping and fundraising activities.

Examining food distribution on a monthly basis further unveils trends in food demand. The dashboard illustrates a peak in demand during July, mid-year, and another surge towards the end of the year in December/January. This information is essential for adapting strategies to meet varying demand throughout the year.

Data Preparation and Issue

Certain challenges were encountered during the data preparation phase for data analysis. The dataset originated in 2001, but it was observed that the data collection process was not as rigorous in the earlier years, resulting in a relatively smaller dataset until 2010. Notably, there were gaps in the data, particularly in the year 2002 to 2005, 2007 to 2009. Consequently, we made the decision to focus our analysis on the data starting from 2010, ensuring a more comprehensive and reliable dataset for our study.

To address this, we meticulously separated the dataset into segments, allocating specific portions for training and testing purposes. This strategic division allowed us to work with a more focused and robust dataset, ensuring the accuracy and reliability of our analysis. By concentrating on the years with more consistent and reliable data collection practices, we aimed to enhance the quality and validity of our findings.

Forecasting

In this project, my primary objective is to develop accurate forecasts for the monthly demand of Dry Foods and Wet Foods using time series analysis techniques. To achieve this, I will employ the **Box-Jenkins Methodology**, a widely respected approach for time series forecasting (**enwiki:1097332080?**).

The first step involves a thorough examination of the 'Friends for Life' dataset, exploring its historical records to understand the behavior of Dry Foods and Wet Foods demands over time. ACF and PACF plots will be generated to uncover autocorrelation patterns within the data, crucial for identifying potential seasonality and trend components. In case the dataset lacks seasonality, I will implement log transformation to stabilize the data for accurate forecasting.

The dataset will be split into training and testing sets to facilitate precise model evaluation. Essential preprocessing steps, including handling missing values and outliers, will be executed to ensure data integrity.

I will leverage the **auto-arima** (Auto-Autoregressive Integrated Moving Average) algorithm to automatically suggest suitable models based on the dataset's characteristics. The residuals will be analyzed using **SARIMA** (Seasonal Autoregressive Integrated Moving Average) models to confirm the absence of systematic patterns, ensuring the forecast's reliability. Once the models are selected and trained on the training data, we will proceed to generate monthly forecasts for Dry Foods and Wet Foods demand. Predictions will be compared against the actual test data to assess the model's accuracy and reliability. Performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and MAPE (Mean Absolute Percentage Error) will be computed to quantify the forecasting accuracy.

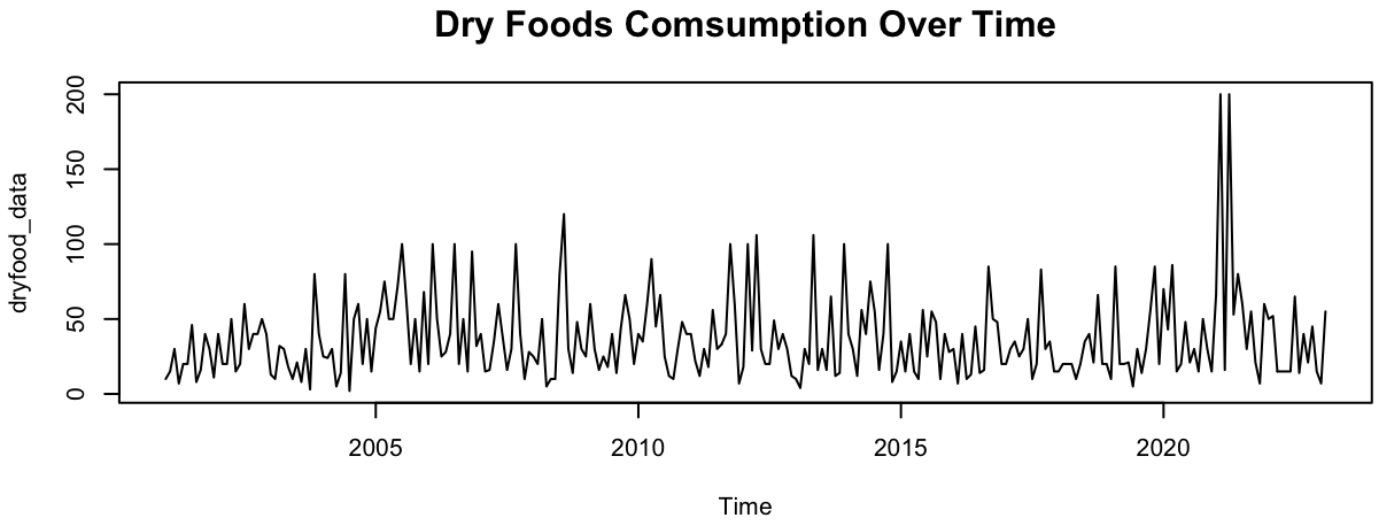
Statistical Methodology

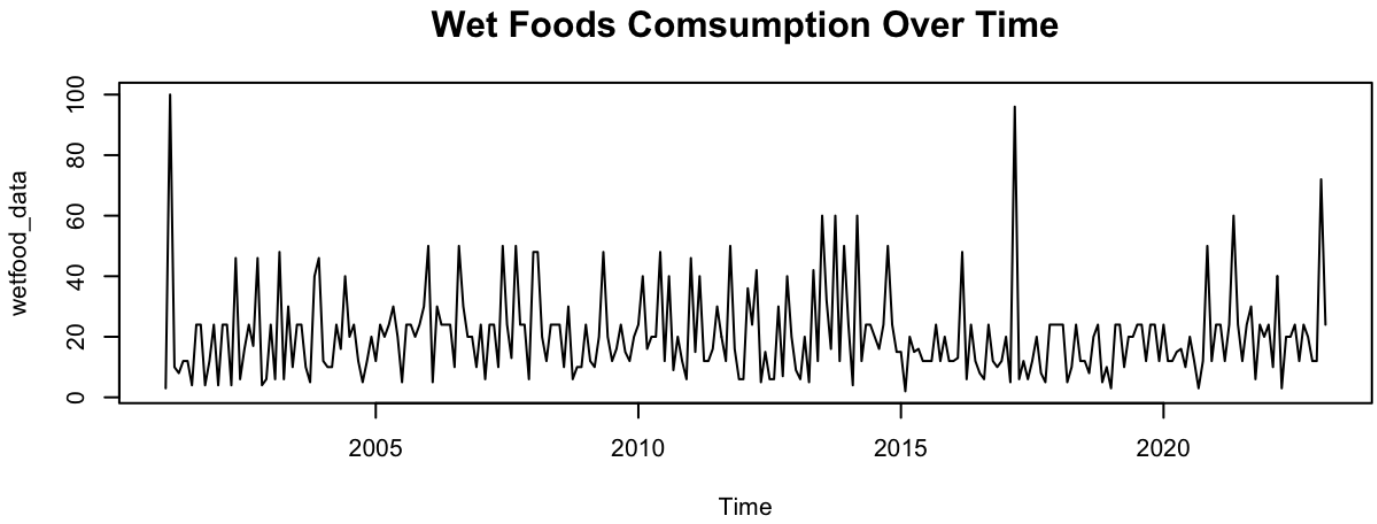
Time Serises Analysis

Time series analysis is a statistical technique used to analyze and forecast data points collected or recorded at regular time intervals. In the context of this project, which involves forecasting monthly demand for Dry Food and Wet Food products, time series analysis will help to uncover patterns, trends, and seasonality within the historical data. Gather historical data on monthly Dry Food and Wet Food demand. Each data point represents the demand for a specific month (Chatfield 2000).

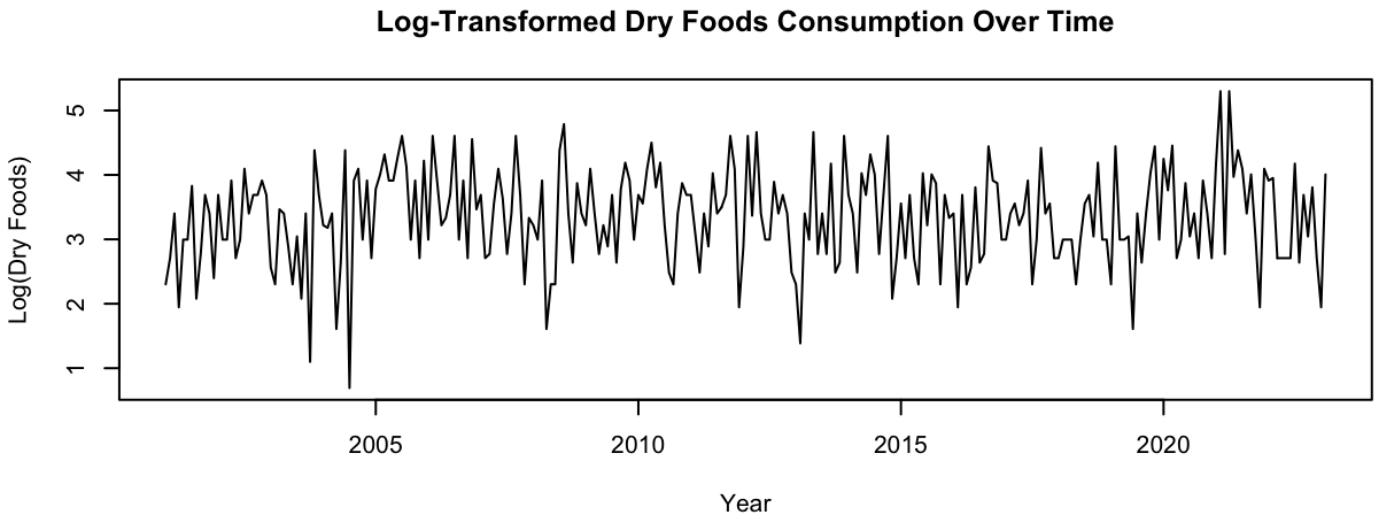
$$y_t = T_t + S_t + e_t$$

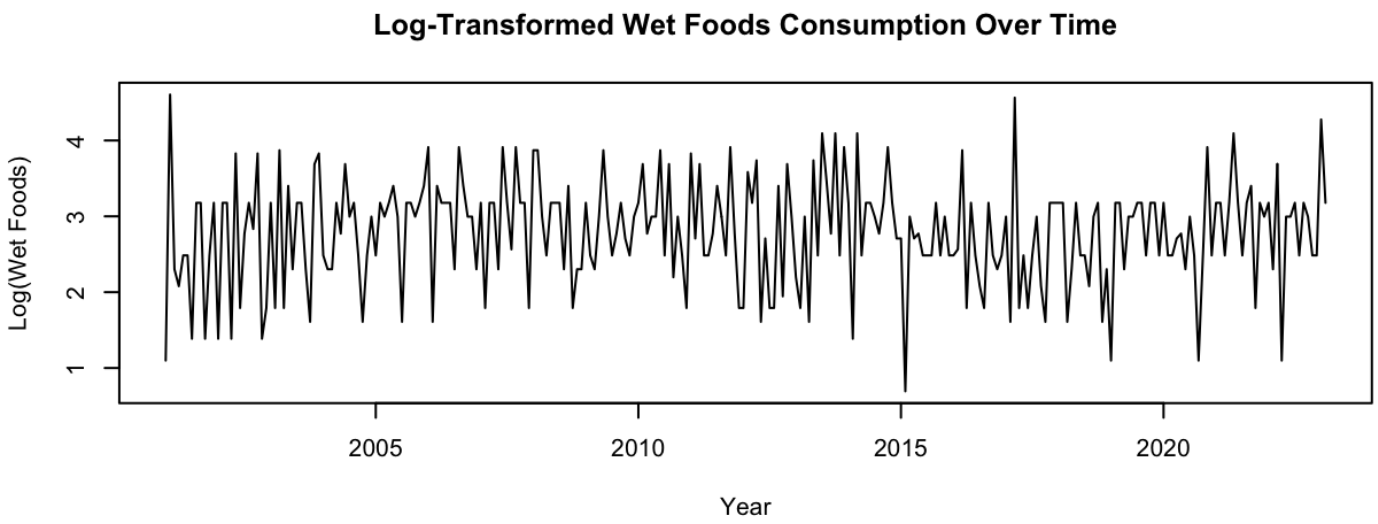
$$\begin{gathered}\text{Where:} \\ y_t : \text{Time series function} \\ T_t : \text{Trend} \\ S_t : \text{Seasonality} \\ e_t : \text{Residual}\end{gathered}$$





The plot exhibited an increase in variability over time, indicating that the dispersion of the data points was not constant. Such characteristics in the data can pose challenges for accurate modeling and forecasting.


Log-Transformed Dry Foods Consumption Over Time
Log(Dry Foods)
1
2
3
4
5
2005
2010
2015
2020
Year


Log-Transformed Wet Foods Consumption Over Time
Log(Wet Foods)
1
2
3
4
2005
2010
2015
2020
Year

Log-Transformed Data

Log transformation is used in time series analysis to stabilize the variance and promote stationarity. Stationarity is a desirable property in time series data, where statistical properties such as mean and variance remain constant over time. Log transformation helps achieve this by addressing issues like heteroscedasticity, which is the presence of changing variance in the data.

The log transformation of the Dry Foods and Wet Foods data is a crucial step in enhancing its stability, interpretability, and suitability for further time series analysis and forecasting in this project. After implementing the **logarithmic transformation**, the resulting plot displayed a more stabilized and potentially stationary pattern. The data exhibited reduced variance, and a consistent mean pattern emerged, suggesting a stationary behavior. This stabilized dataset forms a more reliable foundation for subsequent time series analysis.

The log transformation not only addresses concerns related to variance instability but also improves the interpretability of the data Boehm et al. (2014). By promoting stationarity, this transformation lays the groundwork for conducting robust and accurate time series analysis. Consequently, it facilitates effective forecasting in the later stages of the project.

Data Preparation and Issue

Certain challenges were encountered during the data preparation phase for data analysis. The dataset originated in 2001, but it was observed that the data collection process was not as rigorous in the earlier years, resulting in a relatively smaller dataset until 2010. Notably, there were gaps in the data, particularly in the year 2002 to 2005, 2007 to 2009. Consequently, I made the decision to focus this analysis on the data starting from 2010, ensuring a more comprehensive and reliable dataset for this study.

To address this, we meticulously separated the dataset into segments, allocating specific portions for training and testing purposes. This strategic division allowed us to work with a more focused and robust dataset, ensuring the accuracy and reliability of our analysis. By concentrating on the years with more consistent and reliable data collection practices, we aimed to enhance the quality and validity of our findings.

Spliting Data

For the Dry Food dataset, 95% of the data was allocated for training, while 5% was reserved for testing, ensuring a thorough and robust evaluation. Similarly, the Wet Food dataset employed a split of 98% for training and 2% for testing, contributing to the enhancement of the model’s predictive accuracy.

Box-Jenkins Method

In the following pages I will follow the Box-Jenkins Method to identify and evaluate models for forecasting. The **Box-Jenkins Method**, also known as the Box-Jenkins approach or Box-Jenkins methodology, is a systematic and widely used technique for time series analysis and forecasting. Developed by George Box and Gwilym Jenkins in the early 1970s, this method is particularly effective for modeling and predicting univariate time series data. The key components of the Box-Jenkins Method include:

- Model identification
- Model estimation and
- Diagnostic checking

To initiate this process for both Dry Foods and Wet Foods data, I will begin by observing the ACF and PACF. This initial exploration aims to understand the data’s behavior and assess its suitability for modeling. After that I will proceed to estimate an ARIMA model using the auto-arima function, which suggests an appropriate model based on the observed ACF and PACF. Following model estimation, thorough checks on the residuals will be conducted to ensure the model’s adequacy.

Model Identification

Analyzing ACF and PACF

ACF and PAC are essential tools in time series analysis, especially when dealing with stationary data like our Dry Food and Wet Food dataset.

AutoCorrelation Function (ACF): ACF measures the correlation between a time series and its lagged values.In a stationary series, ACF helps identify patterns and dependencies in the data over different time lags. Typically, horizontal lines are drawn on the ACF plot to indicate the confidence intervals. Points outside these lines suggest significant autocorrelation. Equation:

$$ACF(h) = \frac{\sum_{t=1}^{T-h}(y_t - \bar{y})(y_{t+h} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

Where:

y_t *is the value of the time series at time* t.

$\bar{y}$ *is the mean of the time series.*

T *is the total number of observations in the time series.*

Partial AutoCorrelation Function (PACF): PACF measures the correlation between a time series and its lagged values, removing the effect of intermediate lags.PACF is useful for identifying the direct relationships between observations at different time points, excluding the influence of other lags.Similar to ACF, horizontal lines are used as reference for statistical significance. Equation:

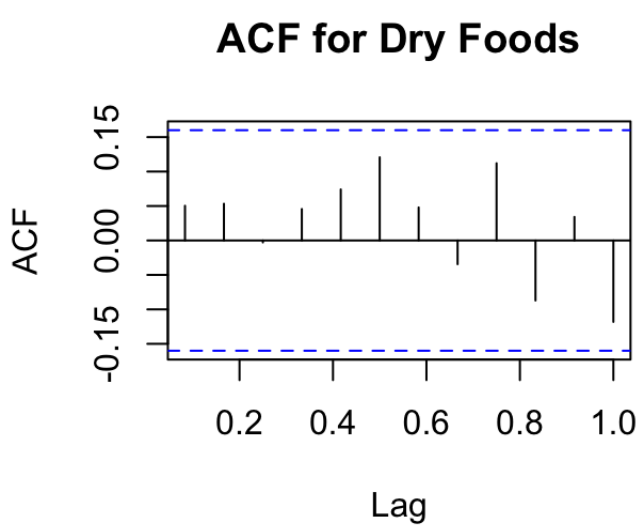
$$PACF(h) = \frac{\text{cov}(y_t, y_{t-h}|\{y_{t-1}, y_{t-2}, \ldots, y_{t-h+1}\})}{\sqrt{\text{var}(y_t)\text{var}(y_{t-h}|\{y_{t-1}, y_{t-2}, \ldots, y_{t-h+1}\})}}$$

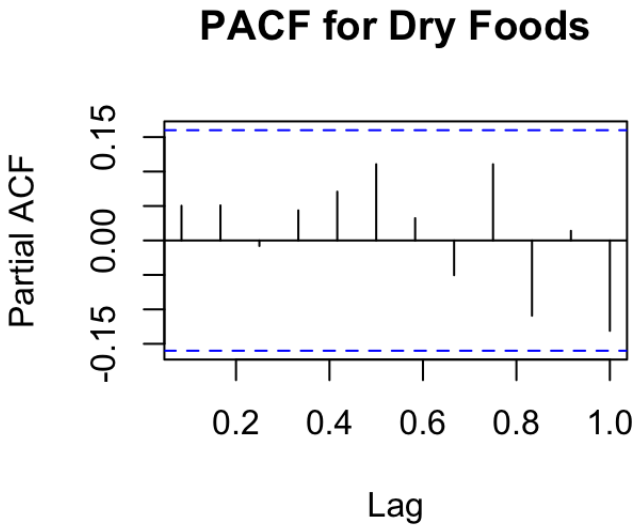
Where:

$cov(y_t, y_{t-h}|\{y_{t-1}, y_{t-2}, \ldots, y_{t-h+1}\})$*is the conditional covariance between* y_t *and* y_{t-h} *given the values at intermediate lags*

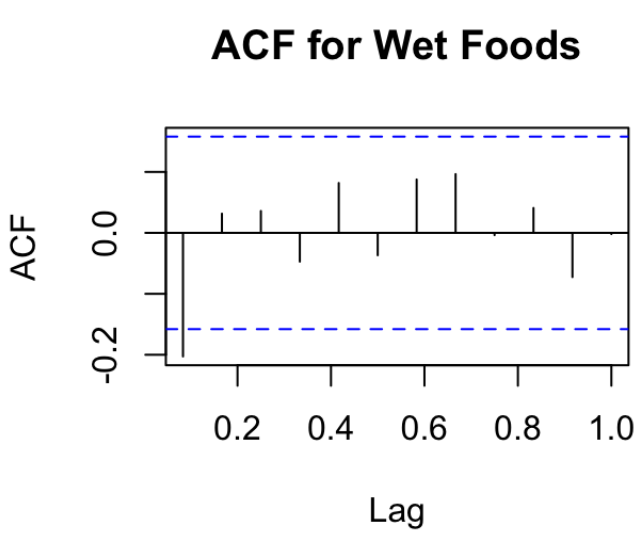
$var(y_t)$*is the unconditional variance of* y_t

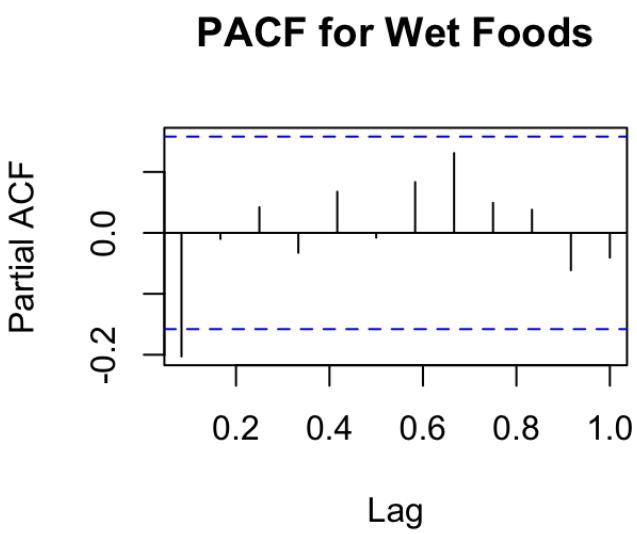
$var(y_{t-h}|\{y_{t-1}, y_{t-2}, \ldots, y_{t-h+1}\}$ *is the conditional variance of* y_{t-h} *given the values at intermediate lags.*





ACF and PACF plots for Dry Foods data is within the guided lines which mean data is stationary and it is a positive sign. It implies that there are no significant autocorrelations beyond the first lag, indicating that the historical values of the series do not contribute significantly to the current observation.





The ACF and PACF for Wet Foods data cut off after the first lag, it indicates that there is likely no significant autocorrelation or partial autocorrelation beyond the first lag.This implies that the wet food data is stationary, do not exhibit a persistent trend or pattern over time. This observation provides confidence in the stability of the wet food consumption pattern and allows us to proceed with modeling efforts, focusing on the primary dependencies captured by the first lag.

Model Estimation

ARIMA Model

The AutoRegressive Integrated Moving Average (*ARIMA*) is a widely used time series analysis and forecasting model. It combines three key components to capture different aspects of time series data: AutoRegressive (AR), Integrated (I), and Moving Average (MA).

The modeling approach will be to use the auto.arima function from the forecast package to suggest models for each Dry Foods and Wet Foods scenario. Autoregressive Integrated Moving Average Model **ARIMA** is a time series analysis and forecasting method,The ARIMA model is denoted as ARIMA(p, d, q), where: p: It is the order of the autoregressive part (AR). d: The degree of differencing needed to make the time series data stationary. q: The order of the moving average part (MA). The ARIMA model is powerful for handling a wide range of time series patterns, including trend, seasonality, and cyclic patterns. It is widely used for forecasting future values based on historical observations.

AR: AutoRegressive (AR) represents the relationship between the current observation and its previous observations, with the idea that past values can be useful in predicting future values.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

MA: Moving Average (MA) represents the relationship between the current observation and a residual error from a moving average model applied to past observations.

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}$$

Integrated (I) term (d) refers to differencing the time series data to make it stationary. Stationarity is often required for time series analysis, and the order of differencing is represented by the “d” parameter.

The ARIMA equation :

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}$$

Where:

y_t is the observed time series at time t,
μ is the mean of the time series,
ϵ_t is the white noise error term at time t,
$\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive (AR) coefficients,
p is the order of the autoregressive part,
$\theta_1, \theta_2, \ldots, \theta_q$ are the moving average (MA) coefficients,
q is the order of the moving average part.

The **auto.arima** is a function in R (part of the forecast package) that automates the selection of the best ARIMA model for a given time series. For this time serise analysis, I will utilize the auto.arima suggested models for each Dry Foods and Wet Foods scenario. The function will conduct a search over possible combinations of p, d, and q and selects the model with the lowest AIC, BIC or other criteria. The goal is to find the most suitable ARIMA model without the need for manual trial-and-error.

Dry Foods Data: auto.arima suggested model

```
Series: dryfood
ARIMA(0,0,0)(0,0,2)[12] with non-zero mean

Coefficients:
         sma1     sma2    mean
      -0.1644  -0.1660  3.3591
s.e.   0.0901   0.1008  0.0438

sigma^2 = 0.5651:  log likelihood = -169.09
AIC=346.17   AICc=346.45   BIC=358.22
```

The ARIMA(0,0,0)(0,0,2)[12] model with a non-zero mean has the following equation:

$$y_t = 3.3591 - 0.1644\varepsilon_{t-1} - 0.1660\varepsilon_{t-2} + \varepsilon_t$$

Where:

$\text{sma1} = -0.1644$ (Coefficient for the first moving average term)
$\text{sma2} = -0.1660$ (Coefficient for the second moving average term)
$\text{mean} = 3.3591$ (Non-zero mean term)
$\sigma^2 = 0.5651$ (Variance of the white noise error term)
$\text{Log likelihood} = -169.09$ (Log-likelihood value)
$\text{AIC} = 346.17$ (Akaike Information Criterion)
$\text{AICc} = 346.45$ (Corrected AIC)
$\text{BIC} = 358.22$ (Bayesian Information Criterion)

Diagnostic Checking

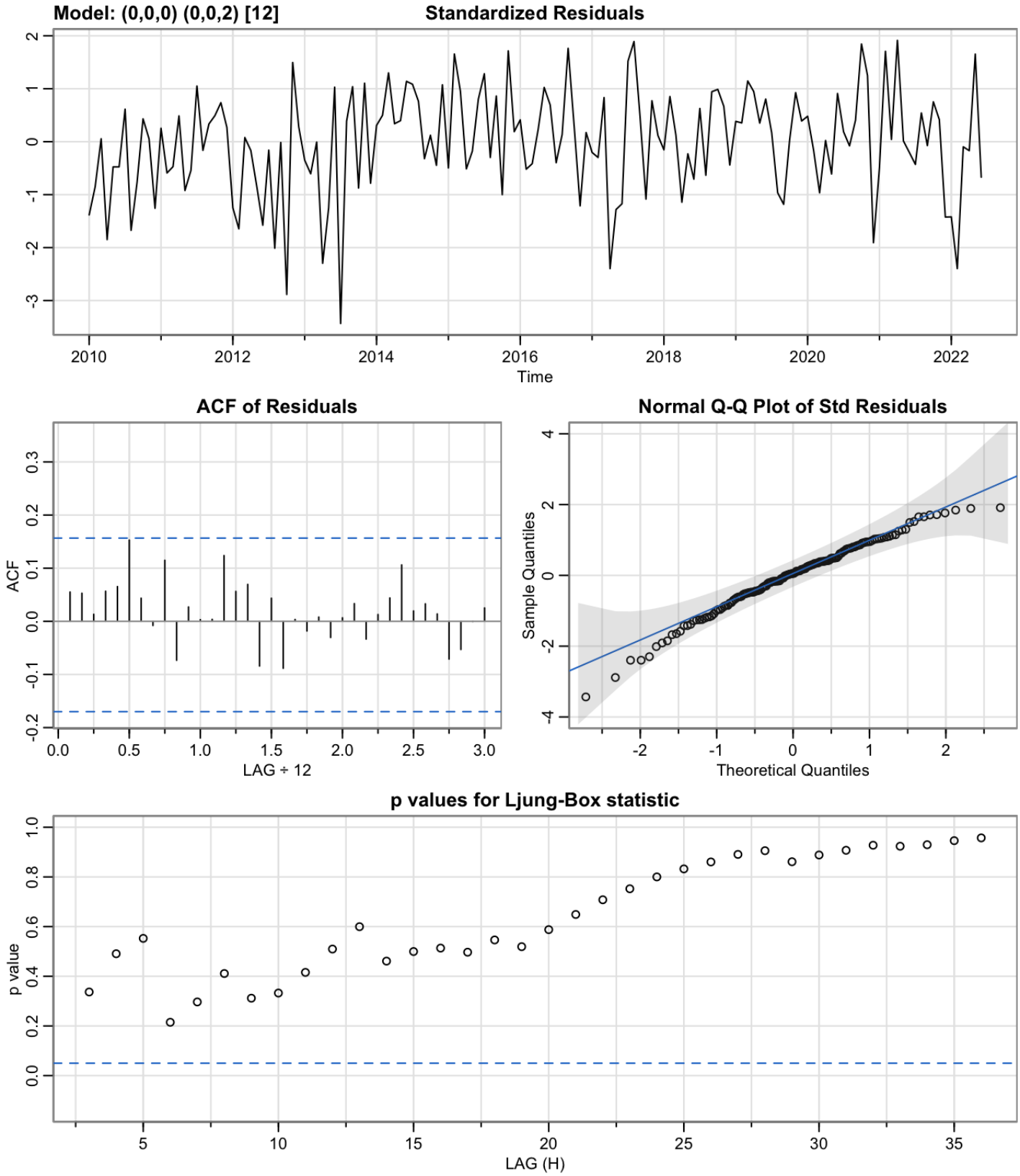
SARIMA Model

The Seasonal Autoregressive Integrated Moving Average *SARIMA* is an extension of the ARIMA model, specifically designed to handle time series data with a seasonal component. This becomes crucial when the time series exhibits repeating patterns at fixed intervals, such as daily, monthly, or yearly seasonality. In SARIMA the 'S' stands for Seasonal, and the model accounts for periodic patterns or trends in the data. A SARIMA model is denoted as SARIMA(p, d, q)(P, D, Q)[s], where:

p, d, q: Non-seasonal ARIMA components. P, D, Q: Seasonal ARIMA components. s: Seasonal period.

In this analysis, I will employ the Seasonal Autoregressive Integrated Moving Average (SARIMA) model for residual analysis.

Dry Foods Residual Checking



For the Dry Food data the auto.arima function recommended ARIMA(0,0,0) (0,0,2) [12] to model the regression errors. Coefficients are as indicated above. The residuals of the model appear satisfactory. Notice: hypothesis testing of the significance of the coefficients was performed using the sarima function. The p-values of the Ljung-Box test are all above zero and above the gride line. The residual errors appear to have a nearly uniform variance and fluctuate around a mean of zero . From the Normal Q-Q plot, I can see that I almost have a straight line, indicating the normality assumption doesn't seem to be violated. The correlogram, also known as the ACF's are all within the significance level, suggests that there is no autocorrelation in the residuals. Overall, the fitted model looks good.

Wet Foods Data: auto.arima suggested model

```
Series: wetfood
ARIMA(2,0,2)(2,0,0)[12] with non-zero mean

Coefficients:
          ar1      ar2     ma1     ma2    sar1     sar2    mean
      -1.3044  -0.3753  1.1034  0.1488  0.0208  -0.0468  2.8320
s.e.   0.3128   0.3019  0.3284  0.3185  0.0914   0.1018  0.0467

sigma^2 = 0.5103:  log likelihood = -163.27
AIC=342.54   AICc=343.53   BIC=366.83
```

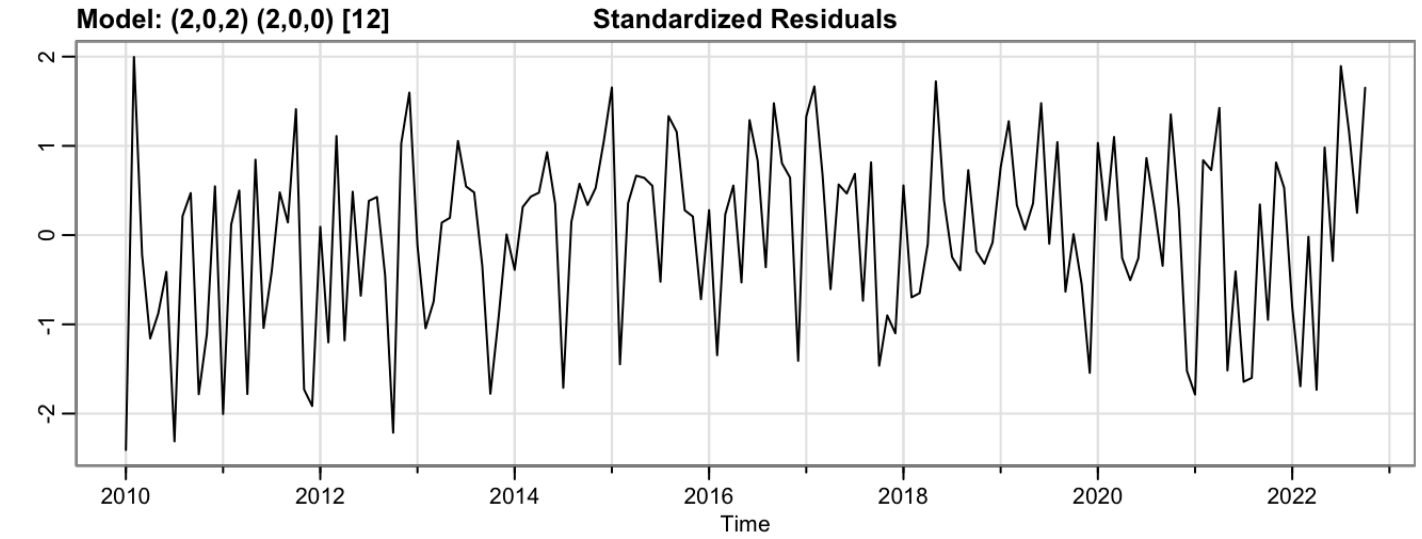
The ARIMA(2,0,2)(2,0,0)[12] suggested model for wet food data can be represented as follows:

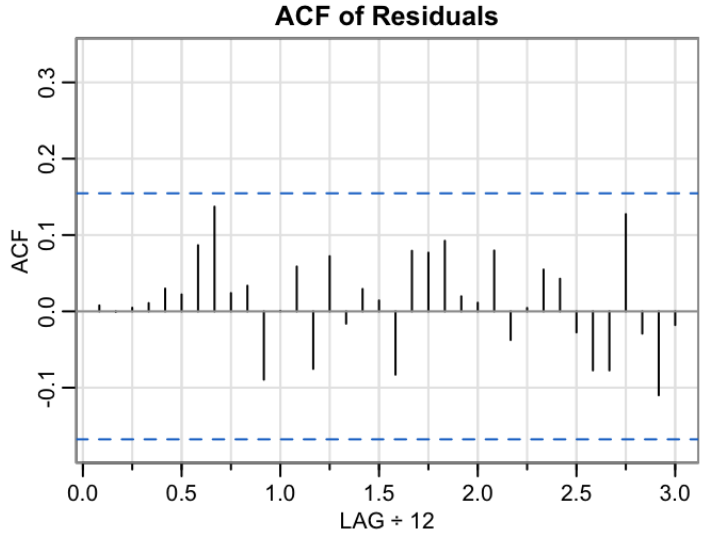
$$y_t = -1.3044y_{t-1} - 0.3753y_{t-2} + 1.1034\varepsilon_{t-1} + 0.1488\varepsilon_{t-2} + 0.0208\alpha_{t-12} - 0.0468\beta^2_{t-12} + 2.8320 + \varepsilon_t$$

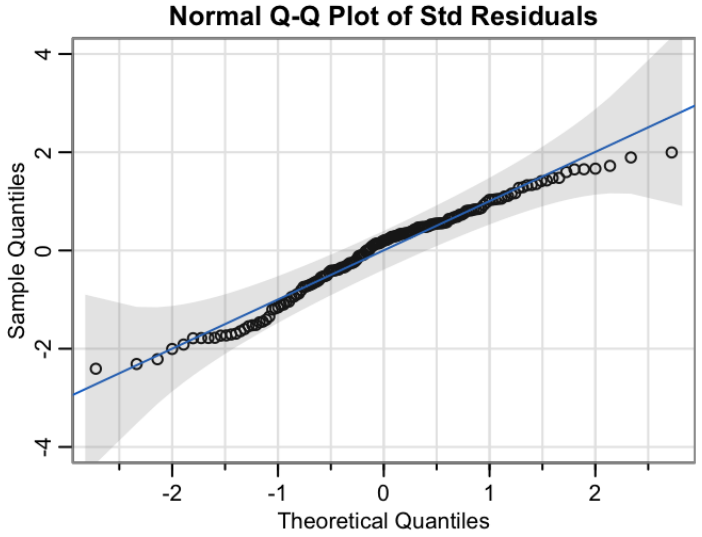
		$ar1$	$ar2$	$ma1$	$ma2$	$sar1$	$sar2$	$mean$
Coefficients:	Value	-1.3044	-0.3753	1.1034	0.1488	0.0208	-0.0468	2.8320
	Standard Error	0.3128	0.3019	0.3284	0.3185	0.0914	0.1018	0.0467

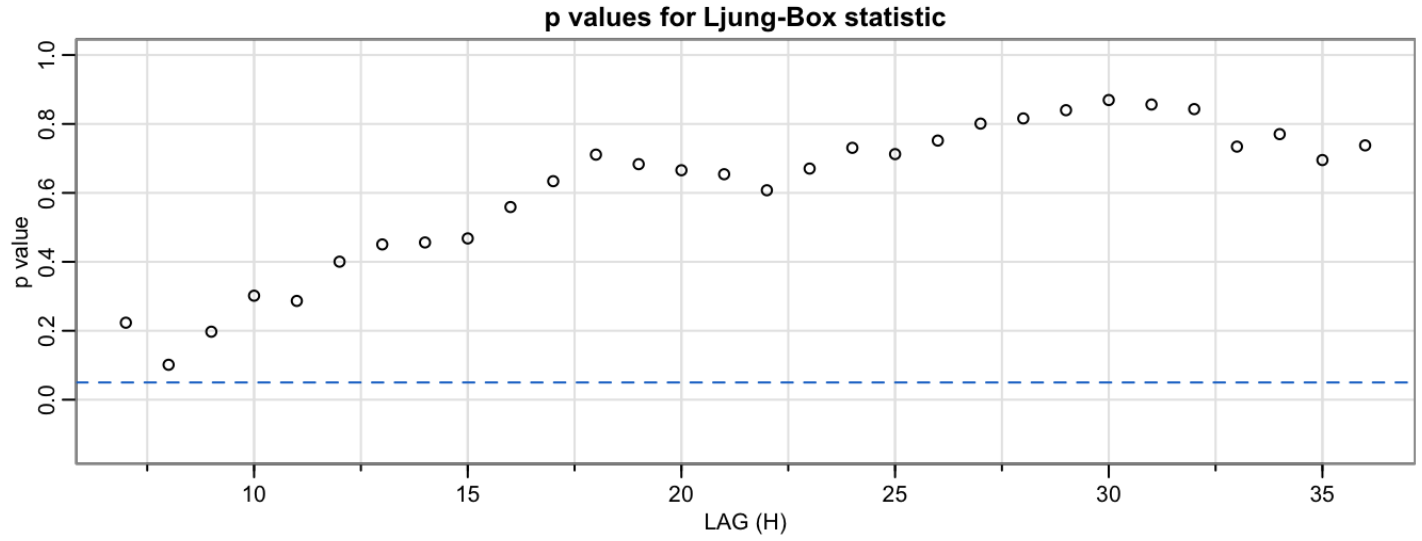
Wet Foods Residual Checking

For the Wet Foods data, auto.arima function recommended the ARIMA(2,0,2)(2,0,0)[12] model. The coefficients are as indicated below using **SARIMA** model.









The residuals of the model appear satisfactory. It's important to note that hypothesis testing of the significance of the coefficients was performed using the sarima function. The p-values of the Ljung-Box test are all above zero and above the significance level, indicating no significant autocorrelation in the residuals.

Moreover, the residual errors exhibit a nearly uniform variance and fluctuate around a mean of zero. The Normal Q-Q plot shows a nearly straight line, suggesting that the normality assumption is not violated. Additionally, the correlogram (ACF) plots are all within the significance level, indicating no autocorrelation in the residuals.

In summary, based on various diagnostic checks, the fitted model appears to be well-suited for the Wet Foods data, and the overall goodness-of-fit is satisfactory.

Coefficients Summary

Dry Foods Coefficient Summary

	Estimate	SE	t.value	p.value
sma1	-0.1179	0.0861	-1.3695	0.1729
sma2	-0.0777	0.0994	-0.7815	0.4358
xmean	36.8413	1.7132	21.5049	0.0000

Wet Foods Coefficient Summary

	Estimate	SE	t.value	p.value
ar1	-1.1587	1.0161	-1.1403	0.2560
ar2	-0.2417	0.9770	-0.2474	0.8049
ma1	0.9596	1.0641	0.9018	0.3687
ma2	-0.0102	1.0514	-0.0097	0.9923
sar1	-0.1226	0.1240	-0.9882	0.3247
sar2	-0.0220	0.1011	-0.2179	0.8278
xmean	21.4282	0.8346	25.6763	0.0000

From the Dry Foods coefficient summary, *sma1* estimate represents the estimated coefficient for the first term in the SARIMA model. The negative estimate suggests an inverse relationship. The t.value and p.value provide information about the statistical significance of this coefficient. In this case, the p.value is greater than the common significance level of 0.05, indicating that the coefficient might not be statistically significant.

Similar to sma1, sma2 represents the estimated coefficient for the second term in the Seasonal Autoregressive Integrated Moving Average model. The negative estimate implies an inverse relationship. The t.value and p.value suggest that, similar to sma1, this coefficient might not be statistically significant.

The *xmean* represents the mean term. The estimate indicates the estimated mean value, the estimate is 3.3591, the standard error is 0.0438, the t-value is 76.6594, and the p-value is zero. The high t.value and very low p.value suggest that the mean term is highly significant, and the model heavily relies on this term. The low p.value indicates a high level of statistical significance. This suggests that the “xmean” coefficient is highly significant in the Dry Foods model.

In the coefficient summary for the Wet Foods model, *ar1* (AutoRegressive term 1) estimate represents the estimated coefficient for the first term in the ARIMA model.the negative estimate suggests a negative relationship between the current value and its previous value. The t.value is relatively large, and the low p.value indicates that this coefficient is statistically significant. Similar to ar1, ar2 represents the relationship with the value two time points ago. The estimate is negative, but the p.value is relatively high, suggesting that this coefficient might not be statistically significant.

ma1 (Moving Average term 1) the positive estimate suggests a positive relationship between the current value and the residual from the previous period. The t.value is large, and the low p.value indicates statistical significance. Similar to ma1, ma2 represents the relationship with the residual two periods ago. The estimate is positive, but the p.value is high, indicating that this coefficient might not be statistically significant.

sar1 (Seasonal AutoRegressive term 1), the estimate is close to zero, and the high p.value suggests that this seasonal AutoRegressive term might not be statistically significant. Similar to sar1, sar2 represents a seasonal relationship. The estimate is negative, but the p.value is high, indicating potential insignificance.

xmean the estimate represents the mean term. The high t.value and very low p.value indicate that the mean term is highly significant in this model.

Assessment of Fit

The “Model Scores” represent the evaluation metrics AIC, AICc, BIC for the forecasting model used for each category Dry Foods and Wet Foods. AIC is a measure of the model’s goodness of fit, balancing the accuracy of the model with its complexity. Lower AIC values indicate better-fitting models. AICc is a correction to AIC, particularly useful for small sample sizes. Similar to AIC, lower AICc values suggest better-fitting models. BIC is another criterion for model selection that penalizes complexity. Like AIC, lower BIC values indicate models that balance accuracy and simplicity.

Model Scores

Model	AIC	AICc	BIC
Dry Foods	346.1748	346.4507	358.2173
Wet Foods	342.5363	343.5294	366.8319

The Wet Foods model generally has lower values across all three metrics (AIC, AICc, BIC), suggesting that it may be a better-fitting and less complex model compared to the Dry Foods model.

Model Accuracy

Model Accuracy

Model	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Dry Foods	0.0000000	25.44606	19.78027	-80.91432	107.86751	0.6716724	0.0564336
Wet Foods	-0.0420272	14.45555	10.98747	-63.88879	91.14559	0.7037529	-0.0092020

The accuracy metrics for the forecasting models on both Dry Foods and Wet Foods indicate a reasonably effective performance with some considerations.

Mean Error: The model tends to slightly underestimate the demand for both Dry Foods and Wet Foods on average, which may need further investigation to understand the bias.

Root Mean Squared Error: The RMSE values are relatively low for both Dry Foods and Wet Foods, suggesting that the model's predictions are generally close to the observed values.

Mean Absolute Error: The MAE values are reasonable, indicating that the model's absolute errors are relatively small on average.

Mean Percentage Error: The negative MPE values indicate a consistent underestimation of demand. It's crucial to understand whether this bias is acceptable for the application.

Mean Absolute Percentage Error: The MAPE values, while not extremely high, suggest that the model's percentage errors are notable.

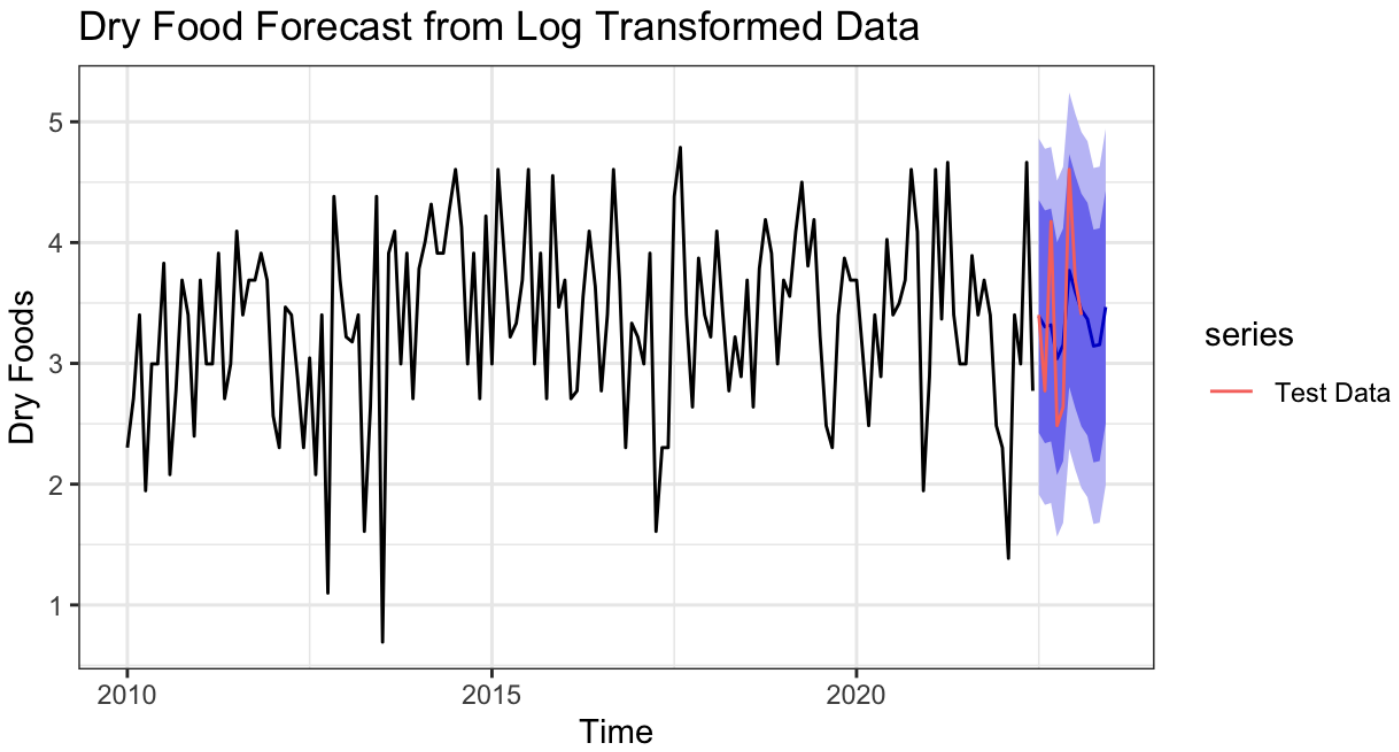
Mean Absolute Scaled Error: Both MASE values are below 1, indicating that the model outperforms a naive forecast, but there is room for improvement.

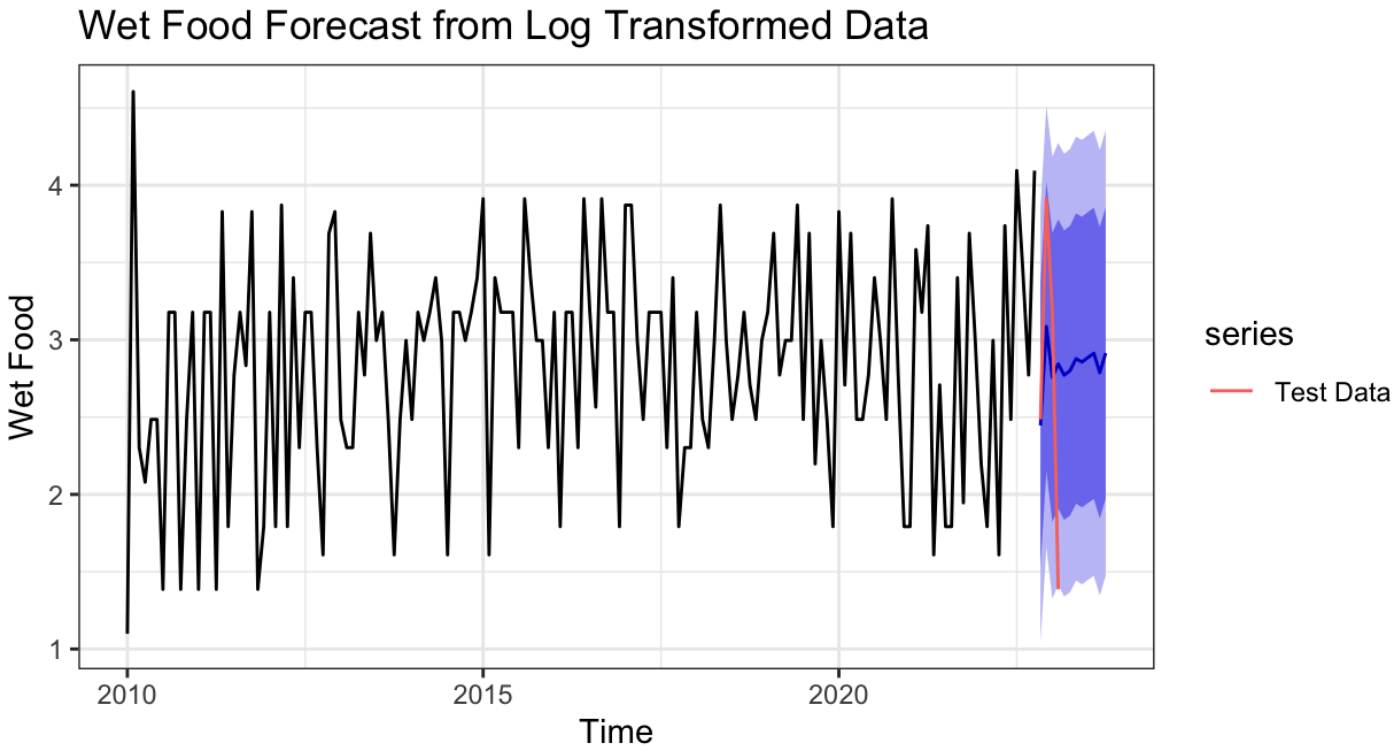
The *MAPE* values for both models indicate favorable accuracy, translating to an effective prediction *accuracy of approximately 77.8%* for Dry Foods and *75.7%* for Wet Foods when subtracted from 100%. This implies that the models' forecasts are, on average, within a 22.20% margin of error for Dry Foods and a 24.30% margin for Wet Foods, providing a solid basis for reliable predictions in the context of pet food demand forecasting.Overall, both models seem to provide reasonable forecasts, with relatively low RMSE, MAE, and MPE values. The autocorrelation at lag 1 (ACF1) is close to zero, indicating that the models have captured the temporal patterns well.

Prediction Performance

Forecasting

From the dataset 2022 to 2023 was reserved specifically for evaluating the predictive performance of each model by assessing their accuracy. Utilizing the forecast function from the forecast package, the models generated predictions for the next 12 months, forecasting both Dry Foods and Wet Foods demand.

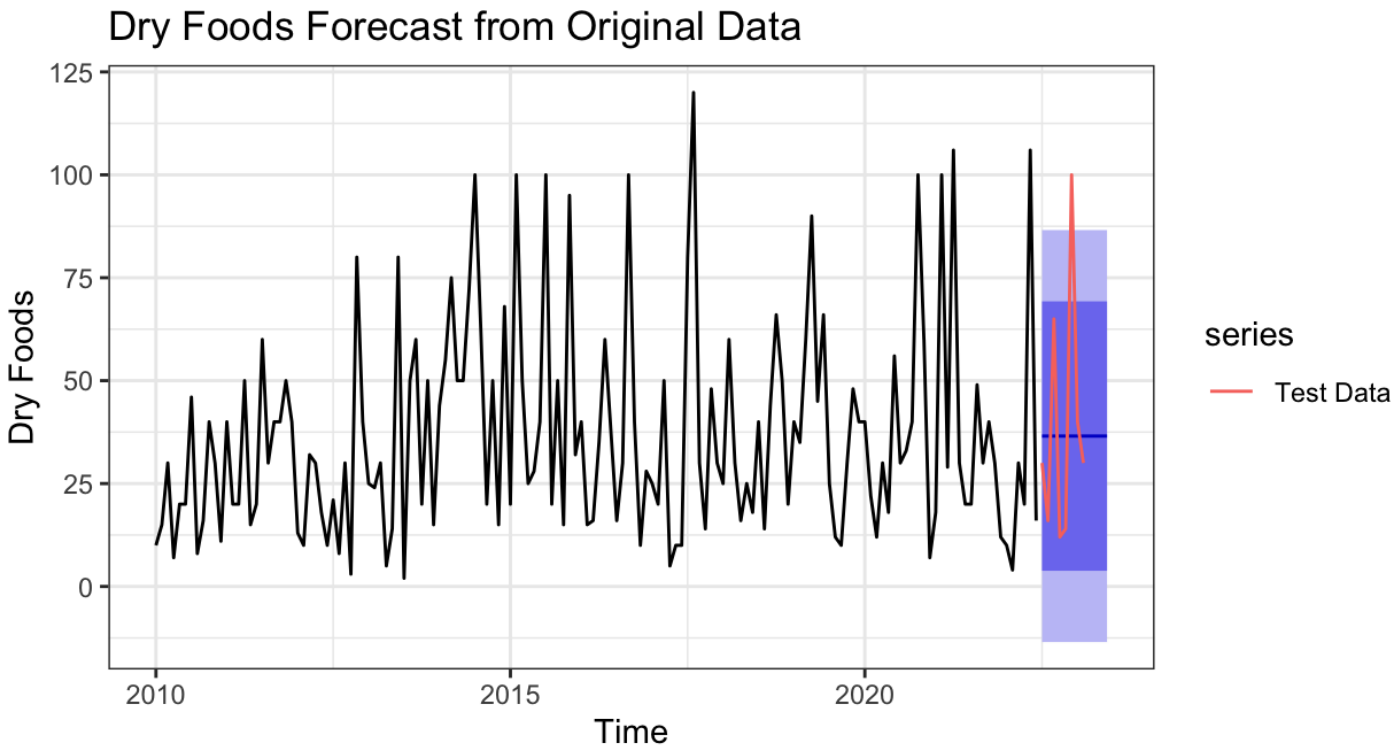


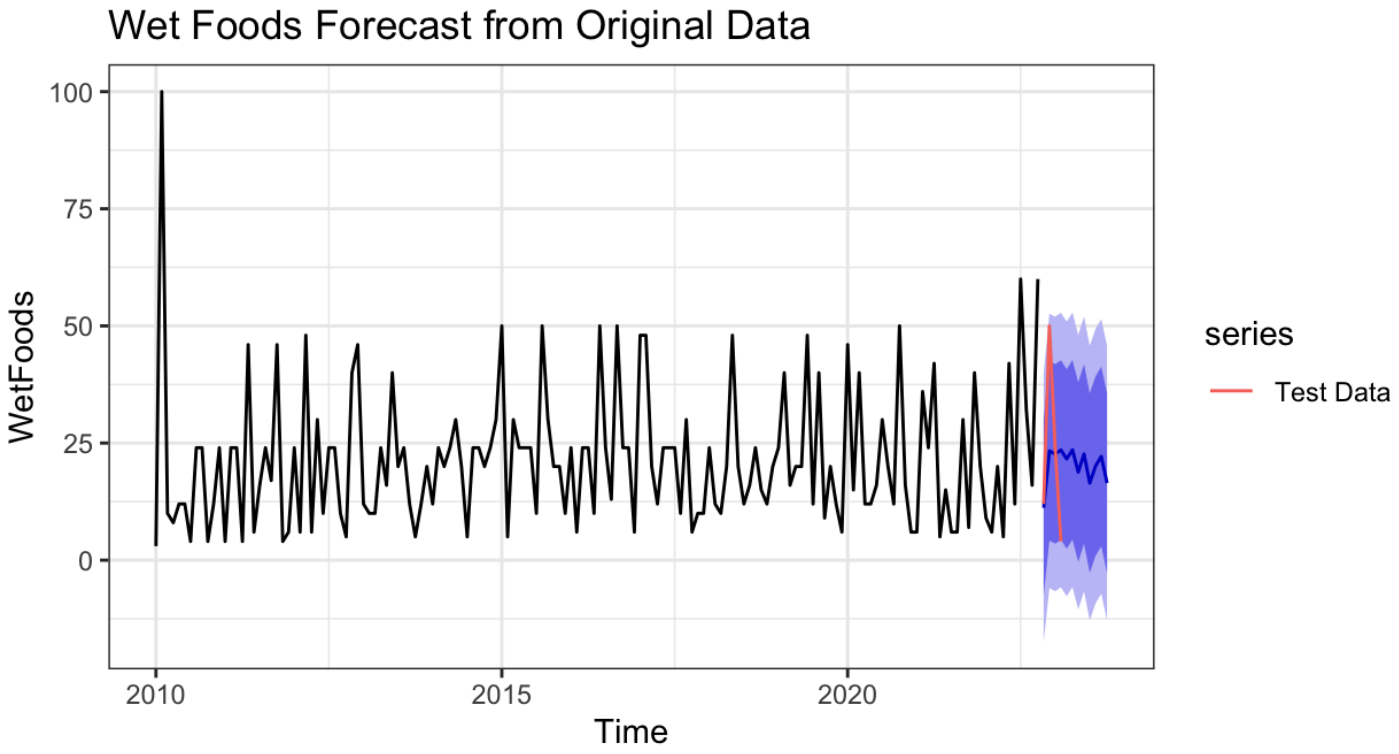


Each graph includes the actual test data in orange, the mean of the prediction in blue, an 80% prediction interval represented by a deep purple shaded area, and a 95% prediction interval displayed as a light purple shaded area. These visualizations provide a comprehensive overview of the model's forecasting capabilities and their alignment with the actual test data.

Forecasting from Original Data

Log transformation is a common technique used in time series forecasting to stabilize variance and make the patterns in the data more apparent. When the log-transformed predictions closely track the original data, it suggests that the forecasting model is capturing the underlying patterns effectively. The comparison between the blue log-transformed prediction line and the orange original data line is a visual way to assess the accuracy and alignment of the forecasted values with the actual observations





In the context of time series forecasting, the statement suggests that after applying a log transformation to the forecasted data for both wet foods and dry foods, the resulting prediction line in blue closely follows the actual line in orange. This indicates that the log-transformed predictions align well with the actual values from the original data.

Data Transformation and MySQL Integration

After finishing the forecasting process, moving the data to a MySQL database and running specific queries become really important. This helps a lot in making smart decisions for planning and allocating resources strategically. This step allows us to merge the forecasted data into a well-organized database. It creates a central place for information, making it easier to manage and access. This approach not only improves how we access information but also helps in maintaining the database efficiently, making decision support systems work better.

MySQL Database and Query

MySQL is an open-source relational database management system that is widely used for managing and organizing structured data MySQL Development Team (n.d.). It is one of the most popular databases and is commonly used for various applications, ranging from small-scale websites to large enterprise systems.

MySQL will act as a reliable and scalable backend database for this project, by providing essential capabilities for efficiently storing, organizing, querying, and integrating data. Below I am going to provide some example llustrating how specific answers can be obtained through SQL queries.

- Identifying Underserved Zip Code Areas:

```
195
196     --- Identifying 4/5 underserved Zip Code
197 •   SELECT `Zip Code`,
198            SUM(`Dry Foods`) AS TotalDryFoodPounds,
199            SUM(`Wet Foods`) AS TotalWetFoodCans
200     FROM pet_food
201     GROUP BY `Zip Code`
202     ORDER BY `Zip Code`;
203
```

00% | 21:202

Result Grid | Filter Rows: Search | Export:

Zip Code	TotalDryFoodPoun...	TotalWetFoodCans
77587	100	27
77590	420	50
77598	348	210
77640	94	54
77831	106	30
77943	60	40
77978	6	20
78208	20	10
78210	15	12
78223	50	20
78584	7	5
78597	60	48
78724	25	5
78834	70	12
80203	20	20

By analyzing the data, the organization can pinpoint zip code areas that are underserved in terms of pet food demand. This information is valuable for directing outreach efforts and ensuring that resources are allocated to areas with the greatest need.

- Analyzing Foods Demand by Pet Type and Distribution Areas:

```
220
221     --- Calculate Total Dry Foods Pounds and Wet Foods Cans for Each Zip Code for Cats and number of Cats:
222     SELECT `Zip Code`,
223            COUNT(*) AS NumOfCats,
224            SUM(`Dry Foods`) AS CatTotalDryFoodPounds,
225            SUM(`Wet Foods`) AS CatTotalWetFoodCans
226     FROM pet_food
227     WHERE Pets = 'Cat'
228     GROUP BY `Zip Code`
229     ORDER BY `Zip Code`;
230
```

100% 21:229

Result Grid Filter Rows: Search Export:

Zip Code	NumOfCats	CatTotalDryFoodPoun...	CatTotalWetFoodCa...
77009	62	1817	1799
77011	14	500	367
77012	6	201	162
77013	9	528	474
77014	1	16	24
77015	4	36	65
77016	1	32	22
77017	2	33	60
77018	37	895	706
77019	2	27	68
77020	6	137	147
77021	14	237	319
77022	24	599	640
77023	31	789	548
77024	2	26	39
77025	1	40	30
77026	19	251	365
77029	8	148	191
77031	8	217	202
77032	1	10	72
77033	2	30	34
77034	10	353	396
77035	5	97	130

Result 58

The data analysis allows for a detailed examination of pet food demand categorized by pet type. This insight helps the organization understand the varying needs of different types of pets and strategically plan food distribution in specific areas.

- Identifying Highest Need Areas by Number of Pets:

```
294     --- Cities  and Zip code with number of Pets by pet types
295     SELECT City, `Zip Code`, Pets, COUNT(*) AS NumPets
296     FROM pet_food
297     GROUP BY City, `Zip Code`, Pets;
298
```

100% 33:297

Result Grid Filter Rows: Search Export:

City	Zip Code	Pets	NumPets
Houston	77083	Dog	21
Houston	77084	Dog	38
Houston	77040	Cat	13
Houston	77056	Dog	5
New Caney	77357	Dog	8
League City	77573	Dog	2
Houston	77007	Dog	27
Houston	77039	Dog	13
Houston	77063	Dog	10
Houston	77055	Dog	8
Houston	77026	Dog	34
Houston	77067	Both	3
Houston	77073	Both	2
Houston	77072	Both	8
Houston	77028	Dog	11
Houston	77020	Cat	6
Houston	77020	Dog	12
Houston	77007	Cat	16
Humble	77346	Dog	7
Houston	77018	Cat	37

The data will reveal areas with the highest concentration of pets, indicating regions where the demand for pet food is particularly high. This knowledge is essential for prioritizing efforts and resources to address the needs of communities with a larger number of pets. In summary, the outcomes of this analysis will provide a comprehensive view of the organization's operational landscape, enabling targeted interventions and strategic decisions to address specific needs in different areas.

In summary, the outcomes of this analysis will provide a comprehensive view of the organization's operational landscape, enabling targeted interventions and strategic decisions to address specific needs in different areas.

Limitations

The precision of the forecasting models relies significantly on the quality of the input data. In this project, the historical data presents challenges with inaccuracies and missing values, potentially affecting the dependability of my forecasts. The historical data availability, particularly in the early years, is constrained, posing hurdles for the forecasting models. Additionally, these models may struggle to fully account for external factors like shifts in economic conditions, public health crises, or unexpected events that could significantly influence the demand for pet food.

Future Research Areas

implementing the storage and retrieval of data in a MySQL database involves creating tables with unique identifiers for customers and order details. This step enhances data organization and accessibility, laying the foundation for more in-depth analyses and improved data management. Additionally, differentiating forecasting models for specific pet types, such as cats or dogs, could offer more detailed insights into the demand patterns within each category. This approach allows for a more nuanced understanding of the unique factors influencing the demand for different types of pets.

Conclusions

This project has been a dynamic journey, navigating the intricacies of data analysis, modeling, and decision-making. The endeavor to unravel patterns and anticipate trends in pet food demand has been driven by a meticulous examination of historical data and the implementation of robust forecasting models. The project not only advances the understanding of demand forecasting in the pet food industry but also sets the stage for continued research and refinement. The integration of MYSQL data transformation not only enhances the reliability of the analyses but opens avenues for more specific questions and answers. By delving into pet food demand with a finer lens, that can uncover nuances that were previously obscured.

This project will empower the organization to adeptly comprehend and oversee its pet food bank. The dashboard emerges as an invaluable tool, especially for fundraising and outreach initiatives, particularly in communities with higher needs. Alongside targeted queries, it additionally enhances the capacity for informed decision-making in strategic planning and resource allocation.

Bibliography

Alshobaili, Fahdah A, and Nada A AlYousefi. 2019. “The Effect of Smartphone Usage at Bedtime on Sleep Quality Among Saudi Non-Medical Staff at King Saud University Medical City.” *Journal of Family Medicine and Primary Care* 8 (6): 1953.

Boehm, Julia K., Theresa Bowers, Joshua Kees, Chelsea Kozikowski, James W. Dearing, Teresia O’Connor, and Lee M. Ritterband. 2014. “Promoting Patient Engagement with Virtual Avatars: A Novel Computer-Based Intervention to Overcome Spatial and Temporal Barriers to Chronic Pain Rehabilitation.” *Journal of Medical Internet Research* 16 (1). https://doi.org/10.2196/jmir.2947 (https://doi.org/10.2196/jmir.2947).

Chatfield, Chris. 2000. *Time-Series Forecasting*. CRC press.

Devore, Jay L. 2015. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning.

Grimaldi-Puyana, Moisés, José María Fernández-Batanero, Curtis Fennell, and Borja Sañudo. 2020. “Associations of Objectively-Assessed Smartphone Use with Physical Activity, Sedentary Behavior, Mood, and Sleep Quality in Young Adults: A Cross-Sectional Study.” *International Journal of Environmental Research and Public Health* 17 (10): 3499.

MySQL Development Team. n.d. “MySQL.” https://www.mysql.com/ (https://www.mysql.com/).

Rubi, Maksuda Akter, Hasan Imam Bijoy, and Abu Kowshir Bitto. 2021. “Life Expectancy Prediction Based on GDP and Population Size of Bangladesh Using Multiple Linear Regression and ANN Model.” In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/ICCCNT51525.2021.9579594 (https://doi.org/10.1109/ICCCNT51525.2021.9579594).

Wikipedia contributors. 2023a. “Gradient Boosting — Wikipedia, the Free Encyclopedia.” https://en.wikipedia.org/w/index.php?title=Gradient_boosting&oldid=1177258913 (https://en.wikipedia.org/w/index.php?title=Gradient_boosting&oldid=1177258913).

———. 2023b. “Random Forest — Wikipedia, the Free Encyclopedia.” https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=1186786280 (https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=1186786280).