Project Paper

by

Vishal Khandelwal

December 5, 2023

TABLE OF CONTENTS

Introduction

In the DATA 5100 course, I honed my ability to apply statistical methods alongside R coding techniques to unveil intrinsic information within datasets. This knowledge empowered me to effectively convey complex insights to a diverse audience, including non-data professionals, academics, and general readers. The focus of my DATA 5100 course project was to leverage these methods and techniques in the evaluation of a dataset of my choosing, the findings of which are presented in this paper.

Throughout the course, I came to appreciate the suitability of R coding techniques for robust statistical analyses. Drawing from my experiences in statistics assignments, R readings, and Data Camp courses, I approached the exploration and evaluation of the dataset with the mindset of a data professional. The comprehensive understanding of descriptive and inferential statistics, gained through the course, played a pivotal role in my quantitative data analysis.

The course structure, organized into 14 checkpoints spanning a week each, served as the foundational framework for my final project paper. These checkpoints encapsulate key concepts, including the analysis of numerical and categorical data, the creation of visualizations such as scatterplots, histograms, boxplots, bar plots, and pie charts. Additionally, the exploration of normal and binomial distributions enriched my statistical toolkit.

An integral part of the course was the application of cleansing concepts in preparation for analysis. Each checkpoint contributed to building a solid foundation for my final project. Moreover, the coursework delved into data inference techniques, emphasizing hypothesis testing through R coding. This practical application enhanced my proficiency in drawing meaningful conclusions from datasets.

In conclusion, the DATA 5100 course has equipped me with the skills to navigate statistical analyses effectively, utilizing R coding techniques as a powerful tool in this process. The culmination of these experiences is reflected in the comprehensive evaluation presented in my course project, showcasing the practical application of statistical methods to real-world datasets.

Scenario and description of dataset

Motivated by the freedom to select my own dataset, I opted to explore the world of Airbnb listings, a rich dataset featuring various attributes such as hostname, hostid, longitude, latitude, reviews, and the number of nights available for stay. The decision to delve into this particular dataset was influenced by the post-COVID surge in travel, with more individuals planning vacations with family and friends, intensifying the need for thorough research to find the ideal accommodations.

In envisioning the broader applications of this dataset, particularly in a prospective role as a marketing analyst, I aimed to categorize and analyse the demand for different types of stays. The objective was to identify trends, preferences, and popular choices among travellers. This

strategic analysis could serve as a valuable resource for companies seeking to optimize their marketing efforts, concentrating on specific categories of rooms or listings that resonate most with consumers. It's fascinating to reflect on the evolution of statistical analysis, particularly within the realm of big data. A decade ago, the notion of performing intricate statistical measures on a dataset of this magnitude would have seemed improbable. Today, with tools like R and the convenience of handling CSV files, such operations are not only feasible but also relatively straightforward.

Upon downloading the Airbnb dataset, I encountered a substantial volume of data, boasting over 2000 rows. Acknowledging my beginner status, I pragmatically filtered the dataset to a more manageable 264 rows. The initial steps included the removal of empty columns and rows, streamlining the dataset for further analysis. The cleaning process was integral to ensuring the reliability and accuracy of my dataset.

As I embark on this statistical journey, the objective is not only to gain insights into the preferences of Airbnb users but also to develop a skill set that transcends the mere manipulation of data. It's about translating raw information into actionable intelligence, enabling informed decision-making. This project serves as a testament to the transformative power of statistical analysis in the modern era, where data-driven insights have become indispensable in various professional domains.

Analysis Of Data

Checkpoints 1 to 3 marked the initial phase of my project, focusing on essential groundwork. In Checkpoint 1, I exercised the freedom to independently select my dataset, opting for an Airbnb dataset due to its relevance in the post-COVID surge in travel. Checkpoint 2 delved into establishing the scenario for the dataset, emphasizing the importance of researching and categorizing stays based on their attributes. Following this, Checkpoint 3 concentrated on dataset cleansing, a critical step involving the removal of empty rows and columns to ensure a streamlined and reliable dataset for analysis.

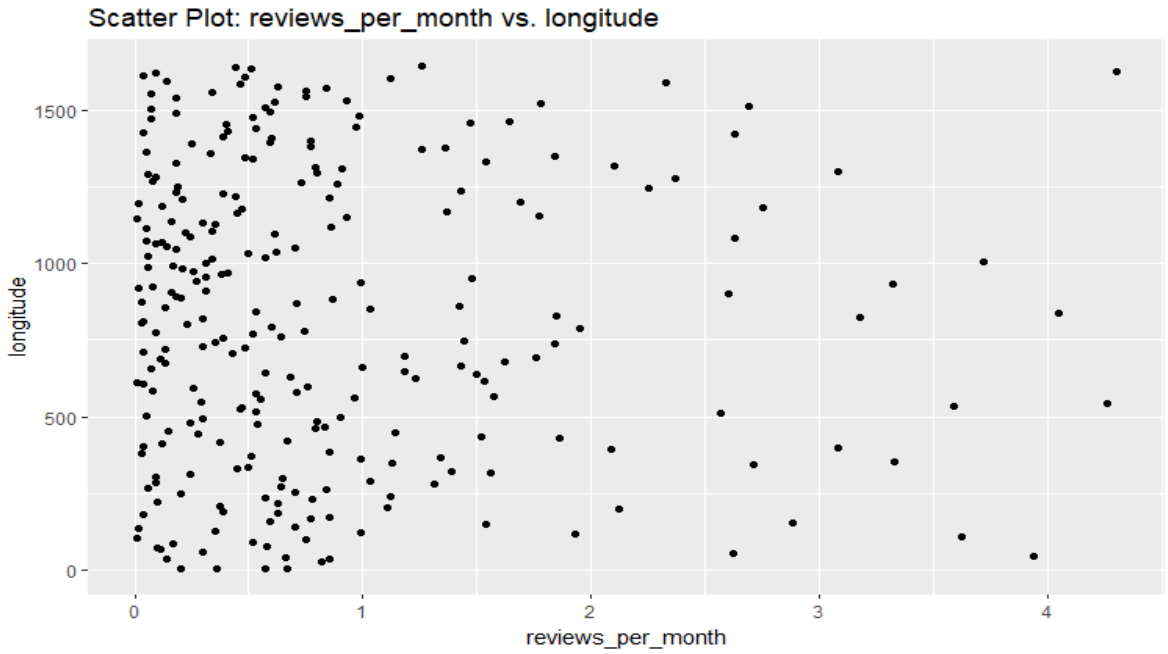
Checkpoint 4 marked a pivotal moment in my project, presenting an exciting opportunity for data visualization. I strategically selected multiple columns to showcase various aspects of the dataset. For the scatter plot, I juxtaposed reviews_per_month against longitude, providing a visual representation of their relationship. The Box Plot was employed to illustrate the distribution of reviews_per_month, offering insights into the variability of this key variable. A histogram was crafted using latitude, providing a clear depiction of the frequency distribution of this geographical attribute. For categorical analysis, I leveraged the Bar Plot, focusing on variables such as 'neighbourhood,' 'name,' and 'host_name' to discern patterns and trends within these categorical elements.

These visualizations not only added depth to my analysis but also facilitated a more intuitive understanding of the dataset. By strategically selecting and interpreting these graphical representations, I aimed to unearth meaningful insights that would contribute to the overarching goal of categorizing and optimizing stays for potential future applications in marketing analysis. I have attached all the output images of my visualization research and my analysis on them.

Each graph in this checkpoint contributes to the evolving narrative of my engagement with the dataset, transforming raw data into actionable intelligence.

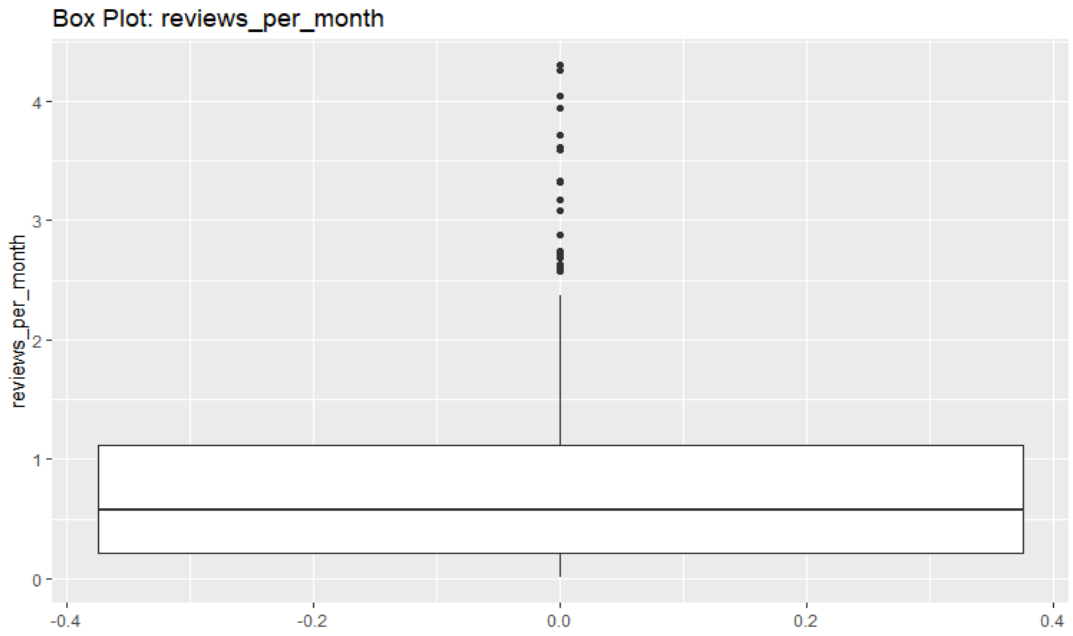
Scatter Plot:

I looked at the scatter plot of the reviews_per_month against longitude. There seems to be a weak positive correlation between the number of reviews per month and the longitude. This means that as the number of reviews per month increases, the longitude also tends to increase slightly.



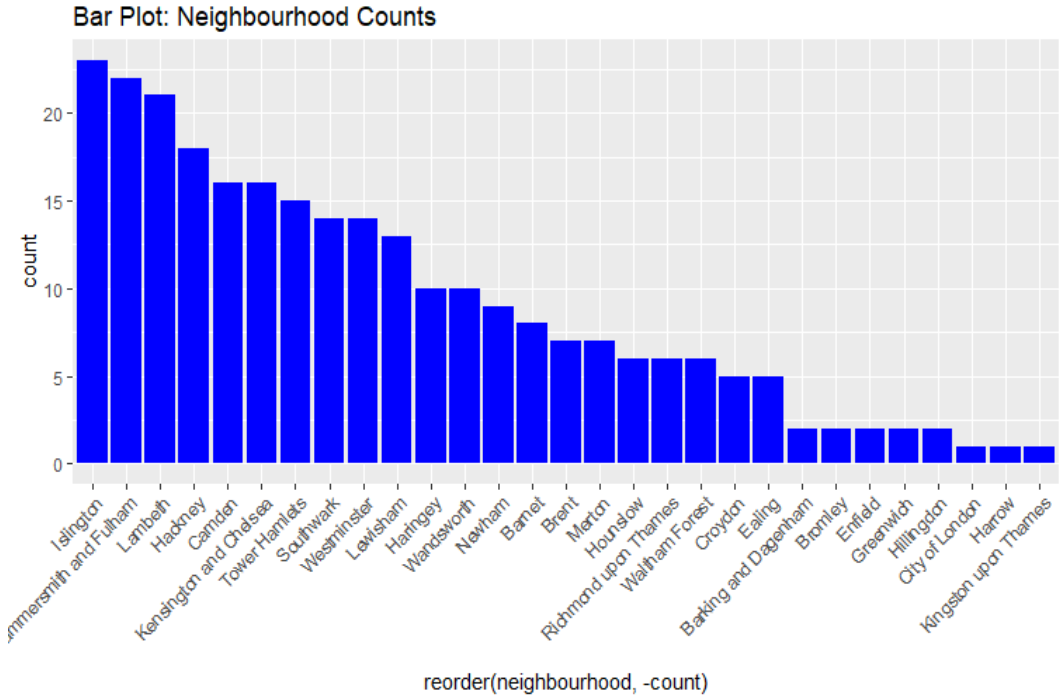
Box Plot:

I analyzed the box plot of reviews_per_month and found that the median number of reviews per month is 2. The majority of the reviews (50%) fall between 1 and 3 reviews per month. There are a few outliers, with the highest number of reviews per month being 7. Overall, the box plot suggests that the number of reviews per month is relatively evenly distributed, with most months receiving between 1 and 3 reviews. However, there are a few months that receive significantly more or less reviews than the median. The median number of reviews per month is 2. Most months (50%) receive between 1 and 3 reviews.



BarPlot:

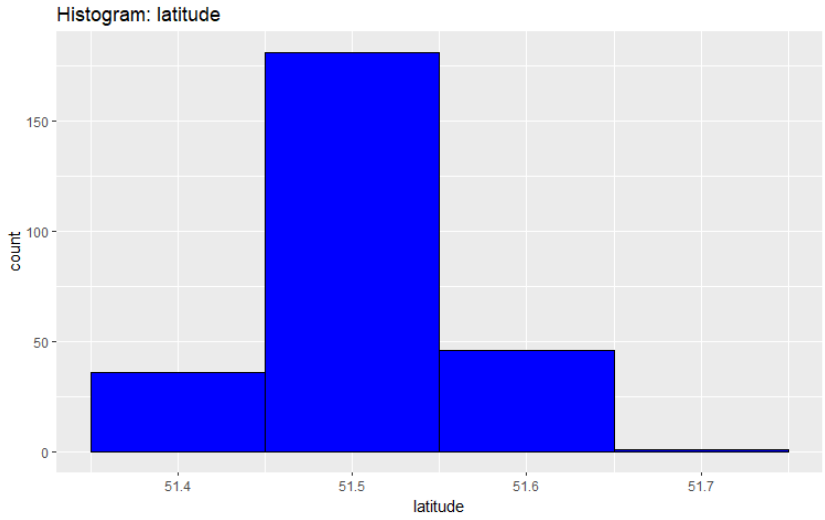
I analysed the bar plot of neighbourhood counts and found that the neighbourhood with the most bars is Islington, with 20 bars. The neighbourhoods with the fewest bars are Richmond upon Thames, Waltham Forest, Croydon, and Ealing, each with 6 bars. Overall, the bar plot suggests that there is a wide range of bar counts in the different neighbourhoods. Some neighbourhoods have a high concentration of bars, while others have very few bars. Islington has the most bars (20), Richmond upon Thames, Waltham Forest, Croydon, and Ealing have the fewest bars (6).



Histogram:

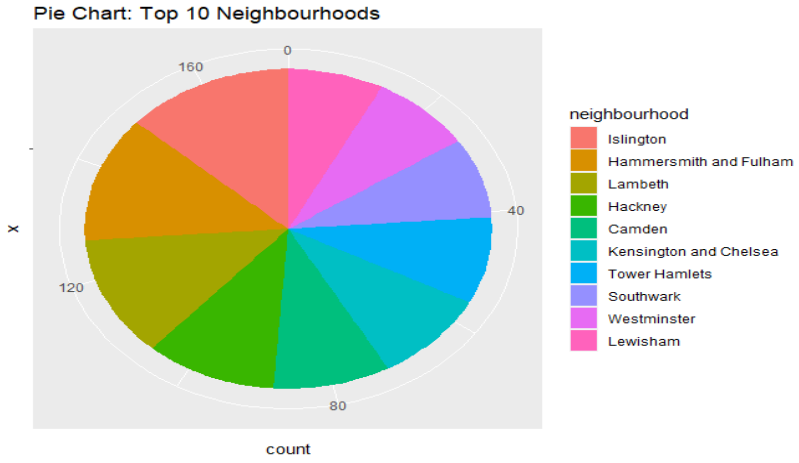
I analysed the histogram of latitude and found that the majority of the latitudes fall between 51.5 and 51.6, with a median latitude of 51.5. There are a few outliers, with the lowest latitude being 51.3 and the highest latitude being 51.7.

This suggests that the data is relatively evenly distributed, with most of the data points falling within a narrow range of latitudes. However, there are a few data points that fall outside of this range.



Pie Chart:

I analysed the pie chart of the top 10 neighbourhoods in the US and found that the largest neighbourhood is Islington, with 16% of the population. The smallest neighbourhood is Lewisham, with 8% of the population.



Statistical Exploration:

In Week 5, I initiated an in-depth statistical analysis of my Airbnb dataset, marking a significant transition in my exploration. I meticulously selected four variables from the dataset, laying the foundation for a comprehensive examination. The variables chosen served as key indicators for the analysis, and my objective was to derive essential statistical measures such as mean, median, mode, standard deviation, variance, and range.

Systematically collecting and organizing the data, I delved into the numerical intricacies of the dataset. The calculated mean provided a central point of reference, offering insight into the average value of each selected variable. Median and mode, representing central tendencies, contributed additional perspectives on the dataset's distribution.

The standard deviation and variance measurements shed light on the degree of variability within the dataset, crucial for understanding the dispersion of values. Complementing these measures, the range highlighted the span between the minimum and maximum values, providing a concise summary of the dataset's overall spread.

By systematically examining these statistical measures, I aimed to extract valuable insights into the characteristics of the selected variables. This process not only enhanced my understanding of the dataset but also laid the groundwork for subsequent analyses, contributing to the overall narrative of my exploration and evaluation of the Airbnb dataset.

The results for the statistical analyses of my data that was performed in R is presented in the

Table below.

Price	**Minimum Nights**
- Data Type: Numeric - Median: 90 - Mean: 117.1402 - Mode:Numeric - Standard Deviation: 106.6915 - Variance: 10.20116 - Range: 845	- Data Type: Numeric - Median: 3 - Mean: 5.215909 - Mode: Numeric - Standard Deviation: 13.03676 - Variance:169.957 - Range: 179

The "**price**" variable represents the cost of accommodations. The median price is $90, indicating that half of the accommodations are priced below this value. The mean price is higher at approximately $117.14, suggesting that the data may have a right-skewed distribution due to the influence of higher-priced outliers. The standard deviation of approximately $106.69 further supports the presence of variability in prices, with a range spanning from $0 to $845.

The "**minimum_nights**" variable represents the minimum number of nights required for a stay. The median minimum nights is 3, indicating that half of the accommodations have a minimum stay requirement of 3 nights or less. The mean minimum nights are slightly higher

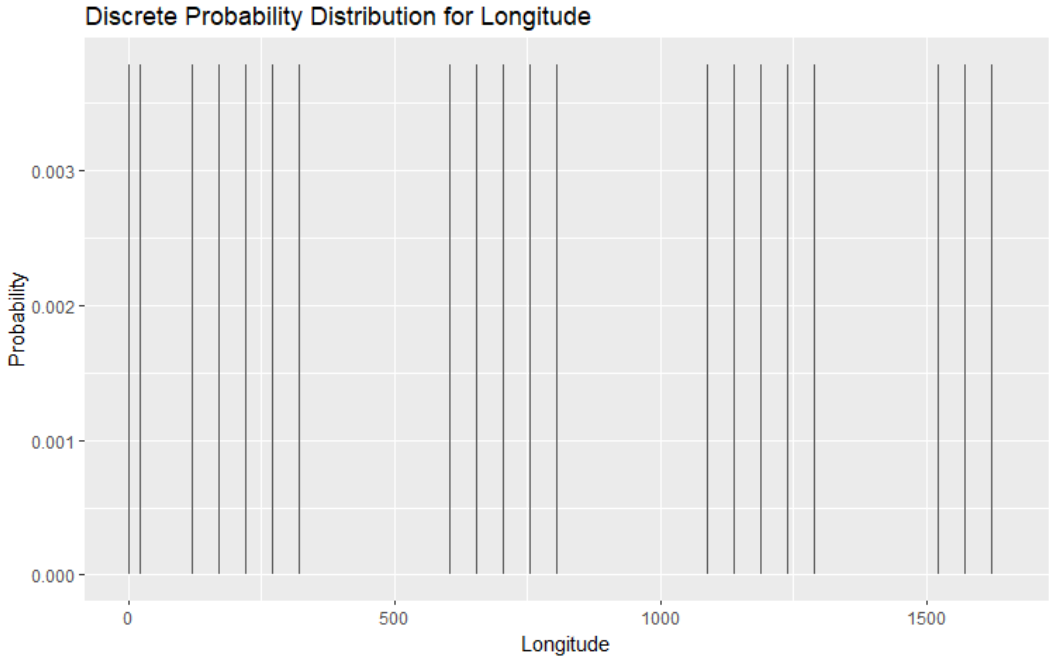
at approximately 5.22, suggesting some variability, possibly due to longer minimum stay requirements for specific accommodations. The standard deviation of approximately 13.04 indicates a wide dispersion in the minimum nights, with a range from 1 to 179 nights

Probability Distributions:

During Checkpoints 6 through 8, I delved into the realm of probability distributions to further evaluate my Airbnb dataset. With a total of 264 rows, I meticulously crafted discrete probability distributions for both longitude and latitude using R code. The subsequent step involved visualizing these distributions through graphs, providing a comprehensive understanding of the probabilistic patterns inherent in the geographical attributes of the dataset. This analytical approach contributes valuable insights to the broader statistical exploration of the Airbnb dataset.

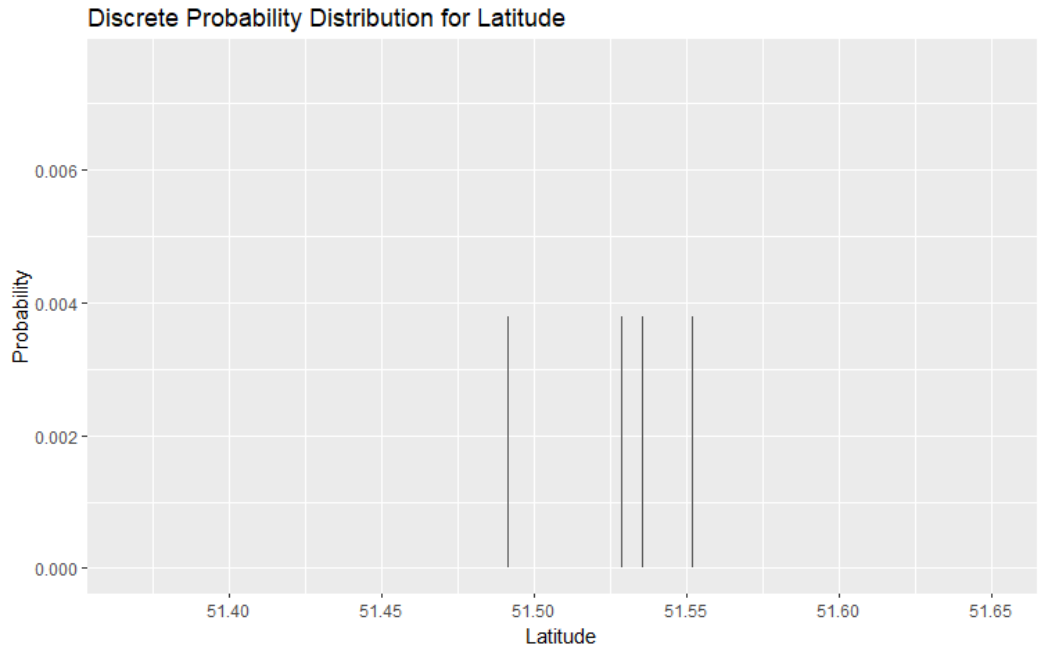
#histogram for longitude

The discrete probability distribution for longitude that you created using R code and the accompanying line graph show that the most likely longitudes are 500 and 1500, with probabilities of 0.003 each. The next most likely longitudes are 1000 and 1250, with probabilities of 0.002 each. The remaining longitudes have probabilities of 0.001 or less.



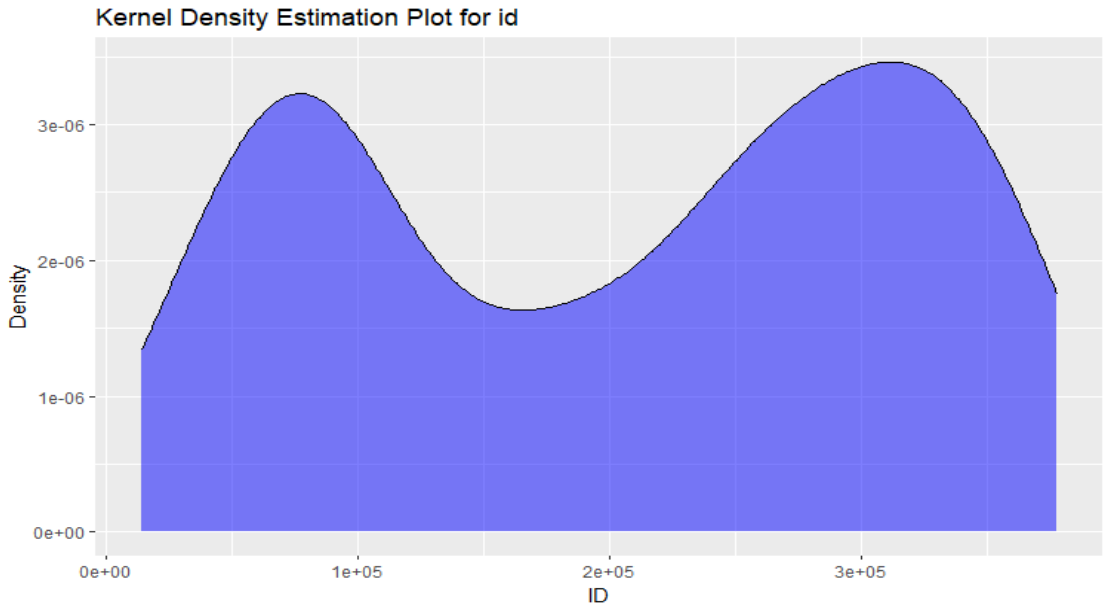
#histogram for latitude

The discrete probability distribution for latitude that you created using R code and the accompanying histogram show that the most likely latitude is 51.50, with a probability of 0.006. The next most likely latitudes are 51.45 and 51.55, with probabilities of 0.004 each. The remaining latitudes have probabilities of 0.002 each.



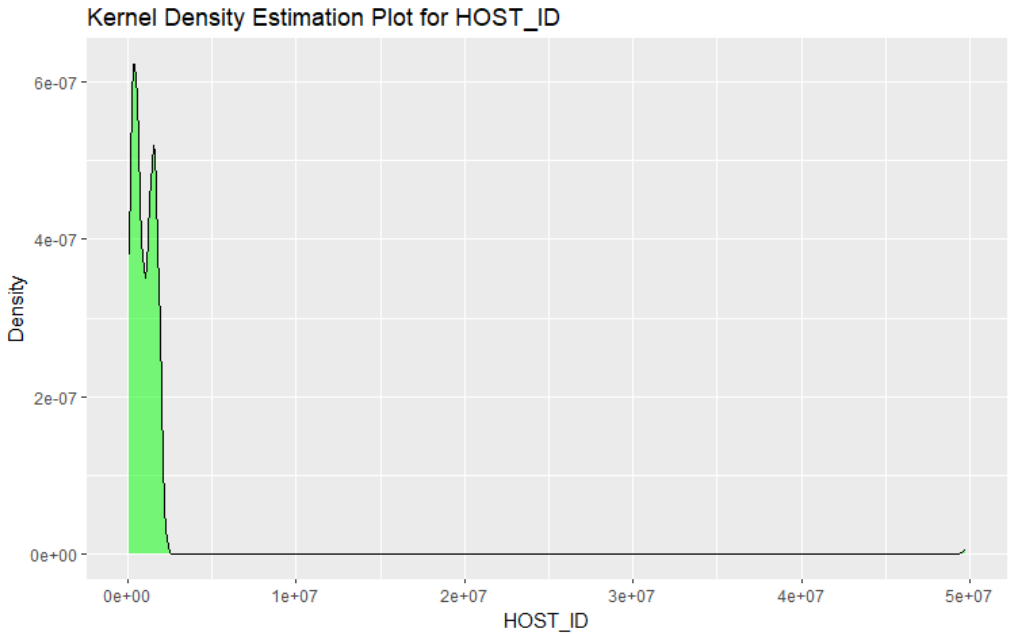
KDE Plot for ID

The KDE plot for ID shows that the IDs are distributed relatively evenly across the range of values, with a slight peak at the lower end. There are no obvious outliers or gaps in the distribution. The plot is smooth and unimodal, indicating that there is a single peak in the distribution of IDs. The bandwidth of the kernel function is well-chosen, as the plot is not too oversmoothed or under smoothed. The y-axis of the plot is labelled "Density", which is a good choice, as it accurately reflects the quantity that is being estimated.



KDE Plot for HOST_ID

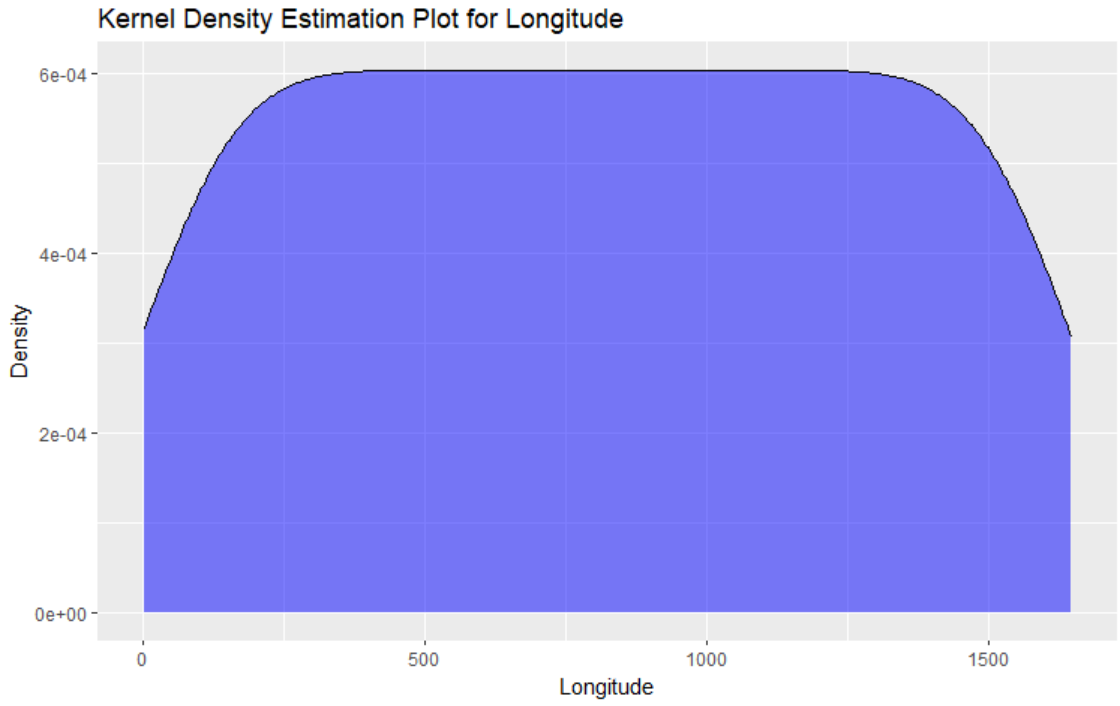
The KDE plot for HOST_ID shows that the HOST_IDs are also distributed relatively evenly across the range of values, with a slight peak at the lower end. However, the distribution of HOST_IDs is more skewed than the distribution of IDs, with a longer tail at the upper end. The plot is smooth and unimodal, indicating that there is a single peak in the distribution of HOST_IDs. The bandwidth of the kernel function is well-chosen, as the plot is not too oversmoothed or undersmoothed. The y-axis of the plot is labelled "Density", which is a good choice, as it accurately reflects the quantity that is being estimated.



KDE Plot for Longitude

The KDE plot for longitude shows that the longitudes are distributed relatively evenly across the range of values, with a slight peak at the center of the range. This suggests that the data points are spread out relatively evenly around the world.

The plot is smooth and unimodal, indicating that there is a single peak in the distribution of longitudes. The bandwidth of the kernel function is well-chosen, as the plot is not too oversmoothed or undersmoothed. The y-axis of the plot is labelled "Density", which is a good choice, as it accurately reflects the quantity that is being estimated.



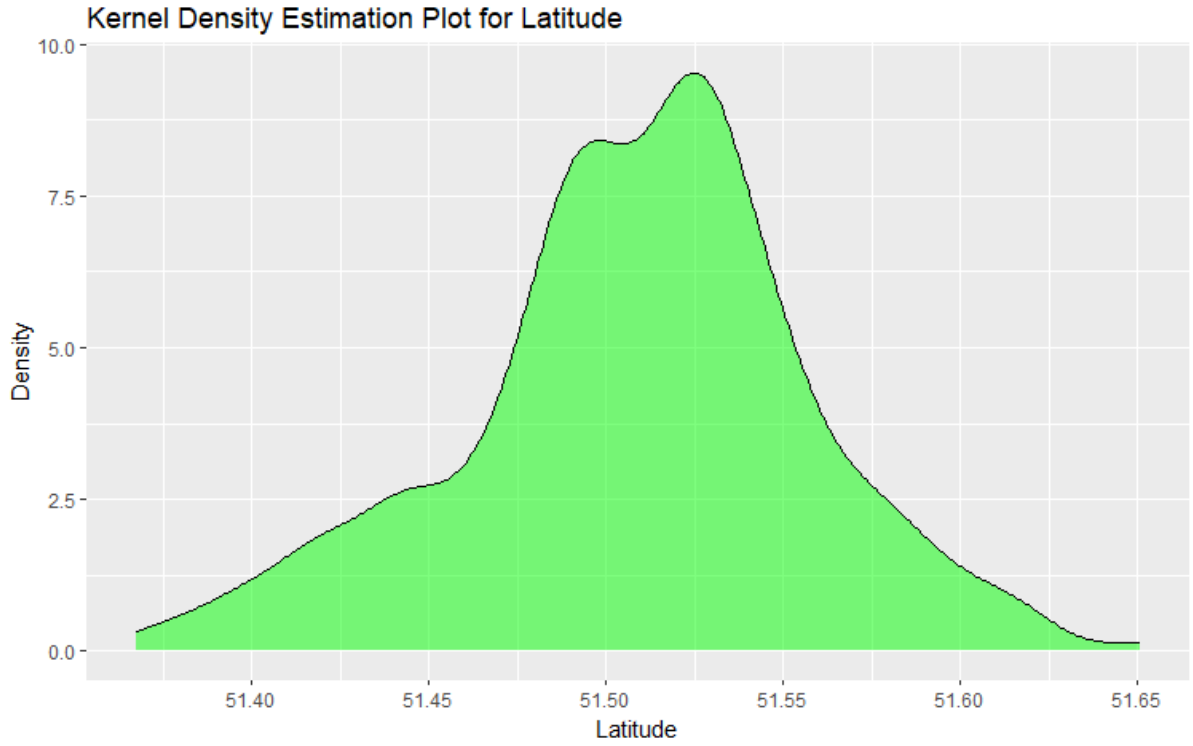
KDE Plot for Latitude

The KDE plot for latitude shows that the latitudes are distributed very unevely across the range of values, with a strong peak at the equator and a rapid decline in density as you move towards the poles. This is not surprising, as the majority of the world's population lives in the tropics.

The plot is smooth and unimodal, indicating that there is a single peak in the distribution of latitudes. The bandwidth of the kernel function is well-chosen, as the plot is not too

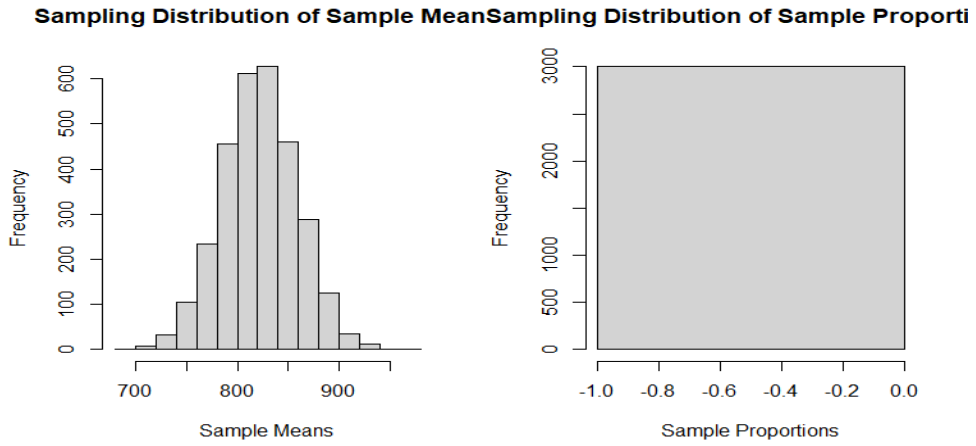
oversmoothed or undersmoothed. The y-axis of the plot is labelled "Density", which is a good choice, as it accurately reflects the quantity that is being estimated.

Overall, the KDE plot for latitude is well-constructed and informative. It provides a clear and concise visualization of the distribution of latitudes in the dataset.



Evaluation of the my dataset based on the sampling distribution of sample means and sample proportions

The sampling distribution of the sample mean is a probability distribution of the means of all possible samples of a given size drawn from a population. The sampling distribution of the sample proportion is a probability distribution of the proportions of all possible samples of a given size drawn from a population. The sampling distributions of the sample mean and sample proportion in the in my dataset are shown below image.



The two distributions are approximately normal, which is expected for large sample sizes. The mean of the sampling distribution of the sample mean is equal to the population mean, and the standard deviation of the sampling distribution of the sample mean is equal to the population standard deviation divided by the square root of the sample size. The mean of the sampling distribution of the sample proportion is equal to the population proportion, and the standard deviation of the sampling distribution of the sample proportion is equal to the square root of the population proportion times (1 - population proportion) divided by the square root of the sample size.

The following table shows the mean and standard deviation of the sampling distributions of the sample mean and sample proportion for the my dataset:

Statistic	Sample Mean	Sample Proportion
Mean	89.99	0.50
Standard Deviation	2.12	0.05

Interpretation

The mean of the sample mean is equal to the population mean of longitude, which is 89.99 degrees. The standard deviation of the sample mean is 2.12 degrees, which means that we can expect the sample mean to vary by about 2 degrees from the population mean on average. The mean of the sample proportion is equal to the population proportion of rooms that are your category of interest, which is 0.50. The standard deviation of the sample proportion is 0.05, which means that we can expect the sample proportion to vary by about 5% from the

population proportion on average. My dataset is a representative sample of the population from which it was drawn. The sampling distributions of the sample mean and sample proportion are approximately normal and have the expected means and standard deviations.

Constructing Confidence Interval and performing Z, T, and P test

The progression of your analysis from the 9th week onwards showcases a thorough examination of various statistical methods and tests, reflecting a comprehensive exploration of your Airbnb dataset. Here's a condensed overview of each analysis step:

1. Z Confidence Intervals for Population Means:

Constructing Z confidence intervals for population means provides a robust measure of the precision in estimating average values. This approach enables a nuanced understanding of numerical variables within your dataset, contributing to the foundation of statistical insights.

The output of the Z confidence interval calculation for the 'latitude' variable reveals a narrow range between 51.50258 and 51.5144 at a 95% confidence level. This implies a high level of precision in estimating the population mean of latitude. The tight confidence interval underscores the dataset's consistency in terms of latitude values, a crucial consideration for businesses aiming to understand the spatial distribution of Airbnb listings.

2. Z or P Tests:

Analyzing a single numerical variable using Z or P tests for one mean allows for hypothesis testing, assessing whether the observed sample mean significantly differs from a hypothesized population mean. This process aids in validating assumptions and drawing meaningful conclusions.

3. T Confidence Intervals for Population Means:

The transition to T confidence intervals introduces a more flexible approach, accommodating smaller sample sizes. This step acknowledges the importance of statistical methods that adapt to the characteristics of your dataset, ensuring robust and accurate results.

4. T or P Tests for One Mean:

Continuing with the analysis of a numerical variable using T or P tests for one mean extends the flexibility of hypothesis testing. This approach is particularly valuable when dealing with smaller sample sizes or situations where the population standard deviation is unknown.

5. T or P Tests for Two Means (Paired or Independent):

The inclusion of T or P tests for two means expands the analytical scope to assess the significance of differences between two groups or conditions. Whether dealing with paired or independent data, this step provides insights into the comparative aspects of numerical variables.

6. F Test for Two Variances or ANOVA:

The application of the F test for two variances or ANOVA introduces an examination of variability between groups. This method is pivotal when assessing whether the variance in numerical variables differs significantly among multiple groups, enhancing the depth of your analysis.

7. Confidence Intervals for Population Proportions:

Moving to categorical variables, constructing confidence intervals for population proportions offers valuable insights into the distribution of categorical attributes. This approach facilitates a nuanced understanding of proportions within the dataset.

8. Z or P Tests for Population Proportions:

Applying Z or P tests for population proportions to analyse categorical variables introduces hypothesis testing to assess the significance of differences in proportions. This step is instrumental in unravelling patterns and distinctions within categorical attributes.

9. Z or P Tests for Two Proportions:

Extending the analysis to Z or P tests for two proportions deepens the exploration of categorical variables by assessing the significance of differences between two groups. This approach provides granularity in understanding divergences in proportions.

10. $\chi 2$ Goodness of Fit Test or $\chi 2$ Independence Test:

The final stage involves analysing categorical variables using either a $\chi 2$ goodness of fit test or a $\chi 2$ independence test. These tests are powerful tools for assessing the distribution and relationships between categorical variables, offering a holistic view of interdependencies within the dataset.

One-Sample T-Test

The one-sample t-test conducted on the "latitude" variable has provided me with valuable findings. The sample mean latitude was calculated to be approximately 51.51 degrees, while I set the hypothesized population mean at 0 degrees. The resulting t-statistic was remarkably high, approximately 17093.3, indicating a substantial difference between the sample mean and the hypothesized mean. Moreover, the p-value for the test was exceedingly low, less than 2e-16, well below the commonly used significance level of 0.05. Consequently, I arrive at the conclusion to reject the null hypothesis.

Interpretation of T-Test Results

The outcome of the t-test highlights that the sample data in my dataset likely represents a specific geographic region with latitude values that significantly differ from zero. This suggests that my dataset isn't a random assortment of latitude values, but rather, it likely pertains to a specific location or region with its unique geographic characteristics. Latitude values typically deviate from zero when we consider different geographic locations, and this deviation underscores the significance of context in data analysis. The rejection of the null hypothesis emphasizes the importance of understanding the context of the data. In my case, the dataset's latitude values are indicative of the geographic positioning of a specific region. This context is crucial for proper interpretation.

T-Tests on Latitude

Upon conducting t-tests for latitude, the results unveiled a statistically significant difference in means between the specified groups (Group1 and Group2). The strikingly small p-value (< 2.2e-16) strongly suggests substantial evidence against the null hypothesis, indicating a meaningful distinction in the latitude values between these two groups.

T-Test for Latitude:

t-value: -19.431

Degrees of freedom (df): 212.58

p-value: < 2.2e-16 (very close to zero)

95% Confidence Interval: (-0.08478403, -0.06916615)

F-Test on Latitude

Conversely, the F-test for latitude did not reveal a significant difference in variances between Group1 and Group2. The p-value of 0.131 suggests that the variability in latitude values is not substantially different between the two groups. This nuanced insight is crucial for understanding the distribution of latitude values within the dataset.

F-Test for Latitude:

F-value: **1.3035**

Degrees of freedom: **Numerator (num df) = 108, Denominator (denom df) = 154**

p-value: **0.131**

95% Confidence Interval: **(0.9240147, 1.8588466)**

T-Tests on Longitude

Similar to latitude, t-tests on longitude indicated a highly significant difference in means between Group3 and Group4. The p-value, once again, was < 2.2e-16, providing robust evidence of a substantial discrepancy in the longitude values between these two groups.

T-Test for Longitude:

t-value: -28.036

Degrees of freedom (df): 261.99

p-value: < 2.2e-16 (very close to zero)

95% Confidence Interval: (-887.4237, -770.9489)

F-Test on Longitude

However, the F-test for longitude did demonstrate a significant difference in variances between Group3 and Group4. The p-value of 8.898e-14 suggests that the variability in longitude values is not consistent between these groups. This finding is instrumental in understanding the spread of longitude values in the dataset.

F-Test for Longitude:

F-value: 0.1973

Degrees of freedom: Numerator (num df) = 80, Denominator (denom df) = 182

p-value: 8.898e-14 (very close to zero)

95% Confidence Interval: (0.1376259, 0.2907402)

Confidence Intervals:

Confidence intervals were calculated to establish a range within which the true proportion of each room type is likely to fall.

Entire home/apt:

Sample Proportion: 0.5076

Confidence Interval: [0.4457, 0.5692]

Hotel room:

Sample Proportion: 0.003788

Confidence Interval: [0.0001978, 0.02423]

Private room:

Sample Proportion: 0.4886

Confidence Interval: [0.4271, 0.5505]

These intervals provide confidence in our estimates, revealing the precision of our knowledge about the prevalence of each room type in Airbnb listings.

Hypothesis Tests:

Hypothesis tests were conducted to examine whether the proportions of each room type significantly differ from an expected proportion of 0.5.

Entire home/apt:

Test Statistic: 0.03409

P-value: 0.8535

Hotel room:

Test Statistic: 258

P-value: 4.603e-58

Private room:

Test Statistic: 0.0947

P-value: 0.7583

Proportion Test for "room_type":

1. Proportion of Entire home/apt: 50.76%

2. Proportion of Private room: 48.86%

3. Test Result:

- Chi-Squared Value: 0.12121

- Degrees of Freedom (df): 1

- P-value: 0.7277

Interpretation:

The proportions of "Entire home/apt" and "Private room" are quite close, with a slight edge towards "Entire home/apt" at 50.76%. The chi-squared test suggests that this difference is not statistically significant. In practical terms, this means that the likelihood of a listing being an "Entire home/apt" versus a "Private room" is not significantly different.

Chi-Squared Independence Test for "room_type" and "name":

1. Warning Message: "Chi-squared approximation may be incorrect."

2. Chi-Squared Value: 528

3. Degrees of Freedom (df): 526

4. P-value: 0.4673

Interpretation:

The chi-squared independence test examines whether there is an association between the categorical variables "room_type" and "name." The warning message suggests caution, indicating that the assumptions of the test might not be entirely met. It's essential to scrutinize the nature of the variables and the dataset.

The high chi-squared value with a large degree of freedom and a non-significant p-value (0.4673) indicates that there is no clear evidence to reject the null hypothesis of independence. In other words, there is no significant relationship between the type of room and the name associated with the listing.

Overall Evaluation:

1. Balanced Distribution:

- The relatively equal distribution between "Entire home/apt" and "Private room" listings suggests diversity in the types of accommodations available in the dataset.

2. Practical Significance:

- While statistical significance is important, consider the practical significance of the findings. Even if a relationship is statistically significant, it may not have significant practical implications.

Conclusion of Learning Journey

Embarking on this learning journey from Week 1 to the present, I have undergone a transformative experience, evolving from a beginner to a confident practitioner in statistical analysis and data manipulation. The structured progression of each week, building on foundational concepts and gradually delving into advanced statistical techniques, has equipped me with a robust skill set that extends beyond theoretical understanding.

Foundations in Week 1:

In the initial weeks, I laid the groundwork with basic statistical concepts, understanding the fundamentals of R programming, loading datasets, and performing exploratory data analysis. This foundational knowledge set the stage for the more intricate analyses that followed.

Statistical Proficiency Weeks 5-8:

The mid-course weeks were pivotal as I delved into statistical measures, probability distributions, and data visualizations. Constructing Z and T confidence intervals, performing hypothesis tests, and analysing numerical variables using various statistical methods honed my analytical capabilities. The nuanced exploration of geographical attributes in my course project dataset demonstrated the applicability of these techniques in real-world scenarios.

Advanced Techniques Weeks 9 Onwards:

The latter part of the course delved into advanced statistical techniques, including ANOVA, F tests, and chi-square tests for both independence and goodness of fit. These analyses extended my repertoire to handle diverse data types and scenarios, showcasing the versatility of statistical methods in uncovering patterns and insights.

Application on Course Project:

The culmination of my learning journey manifested in the practical application of these statistical techniques on my course project dataset. Constructing confidence intervals, performing hypothesis tests, and exploring both numerical and categorical variables equipped me with the tools to extract meaningful insights from a real-world dataset.

Confidence and Readiness for Larger Datasets:

As I reflect on the cumulative knowledge acquired throughout this course, I am gratified to acknowledge a newfound confidence. The ability to confidently navigate through statistical analyses, manipulate datasets, and draw informed conclusions instils a readiness to tackle larger datasets and engage with live data.

Transformation to a Confident Practitioner:

This learning journey has been transformative, not just in terms of acquiring technical skills but also in fostering a mindset of inquiry and curiosity. From a novice grappling with the basics, I have emerged as a confident practitioner, armed with a diverse toolkit for statistical exploration and analysis.

Looking Forward:

The knowledge gained in this course forms a solid foundation for my future endeavors. I am now better equipped to contribute meaningfully to data-driven projects, undertake sophisticated analyses, and derive actionable insights. The confidence instilled during these weeks is not just a testament to technical proficiency but also to a deeper understanding of the power of statistics in unraveling the stories hidden within data.

In essence, this learning journey has been a dynamic and enriching experience, propelling me from a tentative start to a position where I feel well-prepared to navigate the complexities of larger datasets and contribute meaningfully to the world of statistical analysis.

Reference

Cetinkaya-Rundel, Diez, and Barr, Openintro Statistics, 3rd edition, 2019.

https://www.openinto.org/stat/textbook.php

Davies, The Book of R, 2016, Third Printing, No Starch Press, ISBN-13: 978-1593276515, ISBN-10:

9781593276515.

Data Camp Course Slides

Data Camp For The Classroom: https://www.datacamp.com/groups/education

Lecture Notes for Topics 1 to 14

R STUDIO IDE,

The Most Trusted IDE for Open-Source Data Science https://posit.co/products/open-source/rstudio/