

UNIVERSITY OF WESTMINSTER
4DATA001C STATISTICAL MODELING AND ANALYSIS
COURSEWORK
Group 6

	Name	UoW ID	IIT ID
1	Kuruwage Hemindi	W20522533	20221054
2	Yaddehi Lakshitha	W20531555	20221365
3	Ruchintha Dias	W20844558	20232724
4	Muhammad Saneej	W20862255	20220522

Table of Contents

Introduction.....	3
Acknowledgement.....	4
Workload Matrix.....	5
Part 1 : Individual Part [Data Ethics]	
1. Kuruwage Hemindi.....	6
2. Yaddehi Lakshitha.....	8
3. Ruchintha Dias.....	11
4. Muhammad Saneej.....	13
Part 2 : Population for Counties Data Analysis	
2.1.....	16
2.2.....	21
2.3.....	27
2.4.....	36
Part 3 : Motor Trend Car Road Test Data Analysis	
3.1.....	39
3.2.....	41
3.3.....	47
3.4.....	50
3.5.....	58
3.6.....	61

Introduction

This report presents our work for the “DATA001W Statistical Modelling and Analysis” Group coursework at the University of Westminster. The project comprises three main parts: an individual essay on data ethics, data analysis to be conducted in MS EXCEL with population of counties spreadsheet and Data analysis to be conducted in RStudio with “Motor trend car road tests” dataset. The individual essay focuses on the importance of ethical practices in data. The rest of the component involves the analysis of datasets using Excel and RStudio. Our tasks include identifying and cleaning erroneous data, performing statistical analyses, and interpreting results. This practical exercise aims to enhance our ability to apply statistical methods and collaborate effectively as a team. Through this coursework, we aim to deepen our understanding of both ethical and analytical aspects of data science, preparing us for future challenges in the field.

Acknowledgement

We would like to thank Ms. Alqa Husni and Ms. Nirmani for their support and guidance through this module to finish this coursework. We would like to thank our classmates also for helping us to make this coursework successful. And my special gratitude for my group members as their support and dedication has always pushed us to do the work better.

Workload Matrix

Group Member	Task
Kuruwage Hemindi	<ul style="list-style-type: none"> ▪ Data Ethics Essay ▪ Introduction ▪ 2.4 ▪ 3.5
Yaddehi Lakshitha	<ul style="list-style-type: none"> ▪ Data Ethics Essay ▪ Acknowledgment ▪ 2.3 ▪ 3.3 ▪ 3.4
Ruchintha Dias	<ul style="list-style-type: none"> ▪ Data Ethics Essay ▪ Table of content ▪ 2.2 ▪ 3.2
Muhammad Saneej	<ul style="list-style-type: none"> ▪ Data Ethics Essay ▪ Cover Page ▪ 2.1 ▪ 3.1 ▪ 3.6

Part 1 – Individual Part [Data Ethics]

Kuruwage Hemindi

Data Ethics

Data ethics is the principles or theories behind how organizations gather, protect, and use data.

Data gathering is critical to understanding business issues and developing customized services. However, there are also a lot of challenges and risks. Information security is a major issue, businesses with large data sets are more vulnerable to cyber-attacks that result in significant financial losses and damages to their reputation. Ethics, that is what is right and wrong is another important topic. It's difficult to determine when using data is beneficial or when it becomes excessive and begins to observe people too closely.

Concerns regarding individuals profiling abound, particularly around unfairness and biased beliefs. If computer systems are based on biased data or present what people believe to be true, profiling has the potential to strengthen preexisting beliefs. Systems that use profiling to determine creditworthiness or employment eligibility, for instance, may not treat some groups equitably. This may prolong unjust treatment. Furthermore, if people are unaware of how profiling operates, they may come to distrust technology and groups. We question whether it is appropriate to utilize people's information without their consent or an opportunity to correct any inaccurate information because of the unequal distribution of power between the group receiving the data and persons being profiled.

Information leaks have an impact on many individuals and organizations. It may cause people to lose faith in internet sources. People may refrain from utilizing the internet or providing information due to concerns about leaks. This could stifle the expansion of the internet economy and obstruct novel concepts in fields such as online retail, finance, and health. If significant leaks compromise sensitive government information or important systems, they may pose a threat to national security and endanger the public. Leaks are also incredibly expensive, costing billions annually in lost productivity, fines, and increased cyber security requirements. Government, industry, and citizens must all work together to strengthen online safety, establish strict safety protocols, and foster a culture of strength and attention to battle leaks.

A proper understanding of data requires finding a balance between spotting true insights and going on path. Confirmation bias is a major problem when people interpret data to support preexisting beliefs. It's important to be critical, consider alternative viewpoints and ensure that concepts originate from facts rather than external objectives. Being fair also entails understanding the limitations of the facts and refraining from making strong claims with minimal proof. Data from a single group can be erroneous and harmful when extended to the entire population. Fair reading also often considering the implications for individuals and regulations. In order to handle data fairly, one must be truthful, analytical, and concerned about the impact of data on society.

Laws influence the way we handle data. Data ethics depends on the Human Rights Act and other rules in the UK and EU. They defend people's rights to data security and privacy. The Human Rights Act protects people's privacy rights by preventing unauthorized access to their personal information. These regulations support the Data protection Act of 2018 in the UK and General Protection Regulation (GDPR) In the EU. These laws require that data collection be lawful, transparent and restricted to specific objectives. They guarantee that people are in control of their persona information. Additionally, the rules provide mechanism for individuals to access, amend or remove their data, ensuring that others follow the law and are accountable. The laws contribute to the development of a moral framework for data that values persons and promotes technological trust. For this reason, these rules are essential to ensure moral data use in life and encouraging ethical data practices.

To balance technical innovation with respect for individuals' rights and society values, data ethics is crucial.

References

- European Union. (2016). General Data Protection Regulation (GDPR). Retrieved from <https://gdpr.eu>. This document outlines the principles and requirements for data protection under the GDPR.
- UK Government. (1998). Human Rights Act 1998. Retrieved from <https://www.legislation.gov.uk/ukpga/1998/42/contents>. This act provides the legal framework for protecting individual rights in the UK.

Data Ethics

In the digital era, where data has become as asset, the ethical management of data is paramount. Data ethics encompasses the moral principles and practices that guide the collection, usage, dissemination and protection of data.

People must provide informed knowledge and goal to ethical data collecting to take place. Informed consent from persons and exclusive use of data gathered for the intended purpose are prerequisites for ethical data collection. People who give their informed consent are supposed to be fully aware of the data being collected, how it will be used, and the risks of doing so. Purpose limitation makes sure that information is only gathered for specific, acceptable purposes and isn't utilized for unrelated ones without further permission (Florida Taddeo, 2016). For example, in 2021, Clearview AI faced backlash for scraping billions of images from social media without users consent to develop facial recognition technology. By gathering personal information about people without their knowledge or agreement and using it for purposes they didn't agree to, this activity broke in ethical standards (Hill, 2020). Such activities highlight how important it is to use ethical and transparent methods for gathering data.

Utilizing data, profiling involves creating complete personal profiles that predict behaviours or preferences. Although profiling makes customized assistance possible, it also presents serious ethical issues, especially with relation to discrimination and violation of privacy. According to Eubanks (2018), ethical profiling requires protecting people's privacy and ensuring that there are no biases in the process that could result in unfair treatment. The application of system for credit scoring is one example. Large volumes of personal data are analysed by these systems to access trustworthiness.

Financial loss, identify theft and breach of trust are just a few of the serous ethical and everyday consequences that can result from data breaches. It makes moral sense on organisations to put strong security measures in place to safeguard customer information and reduce the like hood of security breaches. According to Culnan and Willims(2009), the consequences of data breaches can go beyond short-term financial losses and include long-term harm to one's reputation and a decline in public confidence . The severe effects of weak data security are best illustrated by the 2017 Equifax breach , which exposed the personal information of over 147 million people. Potential identity theft was a concern for the victims, and Equifax was a hit with financial penalties and serious harm to its brand (Zou and Schaub,2018). This tragedy serves as a reminder of the moral requirement that business prioritize

data security and take precautions against intrusions to preserve public confidence and protect individual's privacy.

To prevent misuse and misinformation, ethical data analysis and distribution are essential. To avoid misunderstandings, information must be presented truthfully, honestly, and within the right context. Incorrect or misrepresented data may result in negative outcomes and poorly informed decisions (Boyd and Crawford, 2012). For instance, data misuse during the COVID-19 pandemic resulted in a large scale spread of false information. There have been instances where incorrect data presentation has been employed to minimize the seriousness of the have ethical implications since they may cause mis understanding among the public, mistrust of the government and negative health issues.

Legal framework like the GDPR and the Human Rights Act are essential for encouraging moral data practices and safeguarding individual rights. To ensure that people's privacy is protected, and they have control over their personal data, the GDPR , for instance, sets tight criteria for data collection , use , and protection (European Union, 2016). Two key principles are the right to erasure, which allows people to ask for the delete of their data, and data reduction, which restricts data gathering to that which is essential. By protecting the right to privacy and making sure that people are protected from unwanted data gathering and misuse, the Human Rights Act also significantly contributes to data ethics (UK Government, 1998).

It is essential to follow ethical guidelines in every aspect of data management since data will continue to play a significant role in society and data ethics will become more and more important.

References

- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662-679.
- European Union. (2016). General Data Protection Regulation (GDPR). Retrieved from <https://gdpr.eu>
- Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, 51(8), 56-59.
- Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.
- UK Government. (1998). Human Rights Act 1998. Retrieved from <https://www.legislation.gov.uk/ukpga/1998/42/contents>
- Zou, Y., & Schaub, F. (2018). Surviving information overload in the big data era. *Informatics*, 5(3), 40.

Data Ethics

In the current digital age, data has become an important asset driving innovation, economic growth and development. However, ethical considerations related to data collection, processing and use has become a critical concern. This is when data ethics matters. Data ethics refers to the moral principles and guidelines that govern the ethics of data collection, processing and use.

Data collection is the foundation of any data-driven system. However, it has several ethical issues. This data can be someone else's privacy. Therefore, they have information privacy. Information privacy refers to the desire of individuals to control or influence data about themselves. Advances in information technology have raised significant concerns about information privacy and its impacts. (Bélanger & Crossler, 2011). The principle of informed consent is also very crucial as it can ensure that individuals are aware of and agree to the data being collected about them. For example, the Facebook-Cambridge Analytica scandal highlighted how data was harvested without users' explicit consent, leading to significant backlash and regulatory scrutiny (Isaak & Hanna, 2018). Ethical data collection practices must prioritize transparency and give individuals control over their personal information.

Profiling involves analyzing data to identify certain characteristics of individuals. although it can offer personalized experiences, it also raises ethical concerns. For example, Algorithms used in recruitment processes may perpetuate unknown biases, leading to discriminatory practices (Raghavan et al., 2020). Ethical considerations necessitate developing fair and unbiased profiling mechanisms that do not disadvantage any group based on race, gender, or socioeconomic status.

Data breaches have a significant risk to individuals' privacy and security. When dealing with sensitive information, it can lead to identity theft, financial loss, and emotional distress. for example Data breach in Facebook in 2018 lead to 50 million users' data. they encountered a security breach that resulted from internal software flaws. Ethical responsibility demands that organizations implement stricter safeguards to protect data and promptly notify affected individuals in the event of a breach.

Ethics are involved in the distribution and interpretation of information, particularly when it comes to accuracy and possible abuse. Data manipulation or misinterpretation can have negative effects by misleading the public and decision-

makers. In order to promote informed decision-making, data results must be presented truthfully, with the proper context and constraints.

Significant influences on data ethics come from the Human Rights Act and the UK/EU legal system. Establishing strict guidelines for data protection, the General Data Protection Regulation (GDPR) in the EU prioritizes people's rights to privacy and data portability (GDPR, 2016). It allows individuals to access and remove their data and requires companies to get explicit consent before processing personal data. Encouraging ethical data handling, the UK Data Protection Act 2018 is in line with GDPR principles.

Private life is protected by the Human Rights Act 1998 (HRA, 1998) under Article 8, which outlines the right to respect for one's private and family life. Through balancing innovation with respect for individual rights, this legislative framework forces organizations to think about the ethical aspects of data processing.

References

- Information privacy refers to the desire of individuals to control or influence data about themselves. Advances in information technology have raised significant concerns about information privacy and its impacts. (Bélanger & Crossler, 2011)
- Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, 51(8), 56-59. doi:10.1109/MC.2018.3191268
- Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469-481. doi:10.1145/3287560.3287591
- General Data Protection Regulation (GDPR). (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council. Retrieved from <https://gdpr-info.eu/>
- Human Rights Act (HRA). (1998). Human Rights Act 1998. Retrieved from <https://www.legislation.gov.uk/ukpga/1998/42/contents>

Data Ethics

Data Ethics and Their Importance

In the modern digital landscape, data is an invaluable resource, driving innovation and decision-making across various sectors. However, the ethical use of data is crucial to protect individual rights and maintain public trust. Data ethics encompasses the principles governing the collection, analysis, dissemination, and use of data, addressing issues such as privacy, consent, and accountability. Legal frameworks, including the Human Rights Act and the GDPR (General Data Protection Regulation) in the UK and EU, provide guidelines and enforce standards to protect individual rights and ensure ethical practices.

Issues Relating to Data Collection One of the primary ethical concerns in data ethics is the collection of data. Collecting data without individuals' explicit consent can lead to significant privacy violations. A prominent example of unethical data collection is the Facebook-Cambridge Analytica scandal, where data from millions of users was harvested without their knowledge and used to influence political campaigns. This case underscores the importance of obtaining informed consent and ensuring transparency in data collection processes (Isaak & Hanna, 2018). Ethical data collection practices require that individuals are informed about what data is being collected, how it will be used, and who will have access to it. The GDPR mandates that data collection must be lawful, fair, and transparent, ensuring that individual privacy rights are protected and that they have control over their personal information (Voigt & Von dem Bussche, 2017).

Informed Consent: Researchers must obtain informed consent from participants before collecting their data. This ensures that individuals understand how their data will be used.

Privacy: Collecting personally identifiable information (PII) without consent can lead to privacy breaches. For example, the Cambridge Analytica scandal involved unauthorized access to Facebook users' data.

Transparency: Organizations should be transparent about data collection practices. Users should know what data is collected and how it will be used.

Individual Profiling Individual profiling involves analyzing personal data to create detailed profiles that can predict behaviors, preferences, and characteristics. While profiling can offer benefits, such as personalized services and targeted marketing, it also raises ethical issues, particularly regarding privacy and discrimination. For instance, predictive policing algorithms have been criticized for perpetuating racial biases, disproportionately targeting minority communities based on historical data patterns (Richardson, Schultz, & Crawford, 2019). To mitigate these risks, ethical profiling practices must be transparent, accountable, and non-discriminatory. The GDPR provides individuals with the right to object to automated decision-making and profiling that significantly affects them, unless there is explicit consent or a legal basis for such activities (European Parliament, 2016).

Algorithmic Bias: Algorithms that discriminate against certain groups (e.g., race, gender) perpetuate societal inequalities.

Targeted Advertising: While personalized ads enhance user experience, they can also manipulate behavior and invade

privacy. Impacts of Data Breaches Data breaches have severe ethical implications, as they often result in unauthorized access and misuse of personal information. The consequences for individuals can include identity theft, financial loss, and emotional distress. A notable example is the Equifax data breach in 2017, which exposed the personal data of approximately 147 million people, highlighting the need for robust data security measures (Swinhoe, 2020). Organizations are ethically obligated to protect the data they collect and store by implementing strong security measures, conducting regular security assessments, and promptly responding to breaches. The GDPR enforces strict data breach notification requirements, mandating that organizations report breaches within 72 hours and take steps to mitigate harm to individuals (European Parliament, 2016).

Security Measures: Organizations must prioritize data security to prevent breaches. Failing to do so harms users and erodes trust. Responsibility: When breaches occur, organizations should promptly notify affected parties and take corrective actions.

Dissemination and Interpretation of Results The ethical dissemination and interpretation of data are crucial to prevent misinformation and bias. Misleading or biased data interpretations can have significant consequences, influencing public opinion and policy decisions. For example, the misuse of statistical data during the Brexit campaign led to widespread misinformation about the economic impacts of leaving the EU (Bennett, 2019). Ethical dissemination requires presenting data accurately and transparently, with clear explanations of methodologies and limitations. Researchers and organizations must strive for objectivity and honesty, avoiding cherry-picking data to support preconceived agendas. The principles of open data and peer review contribute to the credibility and reliability of disseminated results.

Cherry-Picking Results: Selectively reporting positive outcomes while ignoring negative findings misleads stakeholders. Scientific Integrity: Researchers should adhere to ethical guidelines when publishing research results.

The Role of the Human Rights Act and the UK/EU Legal Framework Legal frameworks are fundamental in upholding data ethics. The Human Rights Act 1998 in the UK, which incorporates the European Convention on Human Rights, ensures the right to privacy and provides a basis for data protection. The GDPR further strengthens data protection laws by imposing rigorous requirements on data collection, processing, and security, emphasizing transparency, accountability, and individual rights. These legal frameworks offer mechanisms for individuals to seek redress for data misuse and hold organizations accountable for ethical breaches. By codifying ethical principles into law, these frameworks help foster a culture of responsible data stewardship and protect individuals' rights in the digital age.

References

- Bennett, O. (2019). Misinformation and the EU referendum. *Parliamentary Affairs*, 72(1),204-220.
- European Parliament. (2016). General Data Protection Regulation (GDPR).
- Human Rights Act 1998, c. 42.
- Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and
privacy protection. *Computer*, 51(8), 56-59.
- Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York
University Law Review, 94(2), 192-233.
- Swinhoe, D. (2020). The 15 biggest data breaches of the 21st century. CSO Online.

Part 2 : Population for Counties Data Analysis

In the xl sheet that contain population of counties, there are 4 columns. They are 'County', 'State', 'Ethnicity' and 'Population'. Also, there are 843 records containing data.

2.1

To use data effectively for analysis, it is very important to perform data cleaning. This process involves removing missing entries and correcting logically incorrect entries to ensure that the data are accurate and reliable for analysis.

a) Identifying Missing Entries

To apply conditional formatting to highlight blank cells in a table with a yellow fill, begin by selecting the entire table (A1:D844). Then, go to the "Home" tab on the ribbon and click on "Conditional Formatting". From the dropdown menu, choose "New Rule". In the "New Formatting Rule" dialog box, select "Format only cells that contain." Within this option, set the criteria to "Blanks". Click on the "Format" button, choose yellow as the fill color, and click "OK". Finally, confirm your settings by clicking "OK" again. This will apply a yellow fill to all blank cells in the selected table.

Result_1:

	A	B	C	D
1	County	State	Ethnicity	Population
2	Strafford County	Louisiana	Black African	440,000
3	Sherburne County	Kentucky	Black African	3,000,000
4	Marion County	North Dakota	White	5,200,000
5	Dickinson County	Iowa	Chinese	419,000
6	York County	Minnesota	Black African	4,000,000
7	Wallowa County	Oregon	Chinese	2,200,000
8	Fulton County	Indiana	Chinese	
9	Baraga County	Kansas	Other Black	
10	Iberia Parish	Indiana	Black African	15,950,000
11	Gillespie County	Oklahoma	Other Asian	850,000
12	Cavalier County	New York	Indian	420,000
13	Judith Basin County	Michigan	Black Caribbean	2,300,000
14	Cape Girardeau County	Iowa	Other Black	8,166,666
15	Randolph County	Illinois	Black African	417,000
16	Beauregard Parish	Illinois	Black African	3,000,000
17	Wharton County	South Carolina	Chinese	416,000
18	Pawnee County	Minnesota	Black African	2,000,000
19	Hall County	Texas	Black African	2,300,000
20	Crittenden County	Colorado	Chinese	440,000
21	Giles County	Ohio	Mixed	2,000,000
22	Fulton County	Alabama	Black African	7,750,000
23	Polk County	Michigan	White	
24	Montgomery County	Idaho	Black African	2,000,000

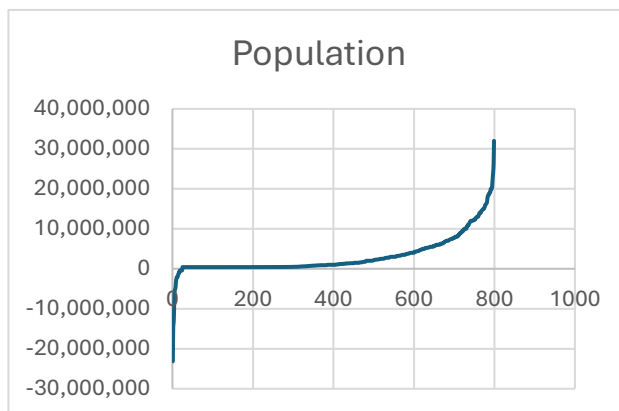
According to the "Result_1" image, the "Population" column contains missing values. Since this data is crucial for our analysis, it should not contain any blank cells. Therefore, it is essential to remove any blank cells from the "Population" column to ensure the accuracy of our analysis.

b) Identifying the erroneous entries

Bad Values - Negative and Zero Values

In the "Population" column, zero or negative values should be removed because a state must have at least one resident and negative population values are not viable. We can figure out the possibility of having erroneous values by using a chart. To do this first copy the 'Population' column's data records (D1:D844) and paste it few columns next to the dataset (G1:G844). Next select copied 'Population' column (G1:G844) and click "Sort & Filter" and select "Sort Smallest to Largest". Now insert this column to the scatter chart.

Result_2:



This graph shows a line starting below -20,000,000, indicating the presence of negative values in the data set. Let's highlight the zero and negative values in the data set.

To highlight cells in the 'Population' columns that are Negative and Zero Values, select these columns (D2:D844), go to the "Home" tab, click on "Conditional Formatting," choose "Highlight Cells Rules," select "Less Than" set the condition to Format cells that are LESS THAN 1 with a Custom Format of Yellow fill and click "OK."

Result_3:

	A	B	C	D
62	Musselshell County	Missouri	Other Black	5,500,000
63	Taylor County	Alaska	White	420,000
64	West Feliciana Parish	Louisiana	Other Black	5,500,000
65	Warren County	Kentucky	Black African	10,000,000
66	Sonoma County	California	Black African	2,300,000
67	Culpeper County	South Carolina	Black African	1,375,000
68	Cass County	North Carolina	Black Caribbean	15,285,714
69	Cerro Gordo County	Colorado	Black African	-5,950,000
70	Morris County	Louisiana	Indian	1,350,000
71	Furnas County	Kentucky	Black African	3,775,000
72	Armstrong County	Pennsylvania	Black African	11,700,000
73	Toombs County	Colorado	White	3,000,000
74	Sullivan County	Oklahoma	Indian	4,000,000
75	Adams County	Illinois	Hispanic	7,611,455
76	Bracken County	Illinois	Black African	5,750,000
77	Lawrence County	Missouri	Chinese	7,100,000
78	Pecos County	North Dakota	Black African	450,000
79	Crane County	North Carolina	Other Asian	20,000,000
80	Smith County	Pennsylvania	Black African	414,000
81	Jasper County	New York	Other Black	4,750,000
82	Woodford County	Arkansas	Other Asian	-23,125,000
83	Stewart County	Ohio	Black African	415,000
84	Livingston Parish	Louisiana	Chinese	440,000
85	Stephens County	California	White	1,000,000

c) Data cleaning - Removing rows with Highlighted cells

To remove highlighted cells, first convert the dataset into a table. Begin by selecting the entire dataset (A1:D844), then go to the "Insert" tab and click on "Table" and confirm by clicking "OK". This will convert the dataset into a table

Result_4:

	A	B	C	D
1	County	State	Ethnicity	Population
2	Strafford County	Louisiana	Black African	440,000
3	Sherburne County	Kentucky	Black African	3,000,000
4	Marion County	North Dakota	White	5,200,000
5	Dickinson County	Iowa	Chinese	419,000
6	York County	Minnesota	Black African	4,000,000
7	Wallowa County	Oregon	Chinese	2,200,000
8	Fulton County	Indiana	Chinese	
9	Baraga County	Kansas	Other Black	
10	Iberia Parish	Indiana	Black African	15,950,000
11	Gillespie County	Oklahoma	Other Asian	850,000
12	Cavalier County	New York	Indian	420,000
13	Judith Basin County	Michigan	Black Caribbean	2,300,000
14	Cape Girardeau County	Iowa	Other Black	8,166,666
15	Randolph County	Illinois	Black African	417,000
16	Beauregard Parish	Illinois	Black African	3,000,000
17	Wharton County	South Carolina	Chinese	416,000
18	Pawnee County	Minnesota	Black African	2,000,000
19	Hall County	Texas	Black African	2,300,000
20	Crittenden County	Colorado	Chinese	440,000
21	Giles County	Ohio	Mixed	2,000,000
22	Fulton County	Alabama	Black African	7,750,000
23	Polk County	Michigan	White	
24	Montgomery County	Idaho	Black African	2,000,000

To display only the rows with highlighted cells, click the white box on the right side of the 'Population' column header and select "Filter by Color" and choose the yellow color box.

Result_5:

	A	B	C	D
1	County	State	Ethnicity	Population
8	Fulton County	Indiana	Chinese	
9	Baraga County	Kansas	Other Black	
23	Polk County	Michigan	White	
30	Knox County	South Carolina	Black African	
34	Fisher County	Pennsylvania	Black Caribbean	
69	Cerro Gordo County	Colorado	Black African	-5,950,000
82	Woodford County	Arkansas	Other Asian	-23,125,000
87	Lee County	South Carolina	Black African	
91	Fannin County	Pennsylvania	Hispanic	
96	Wabash County	Colorado	Black African	
107	Tillamook County	New York	Black African	-420,000
117	Cottle County	Arkansas	Black Caribbean	
154	Worth County	Idaho	Black African	
178	Hardin County	Tennessee	Black African	
183	Meigs County	Ohio	Black African	
187	Haywood County	Pennsylvania	Black African	
191	Monroe County	Missouri	Other Asian	
193	Mitchell County	Illinois	Other Asian	
197	Winnebago County	Indiana	Black African	
201	Richmond city	South Dakota	Chinese	-426,500
208	Nelson County	Texas	White	
209	Ochiltree County	South Dakota	Black African	
219	Wallowa County	Nebraska	Chinese	-10,500,000

Now clear the rows containing highlighted cells by selecting the entire rows except the first row (A8 – D832), right-click and choose "Delete" and click "Entire Sheet Row". Afterwards to bring back the rest of the data records, click the white box on the right side of the 'Contract end' column header, select "Filter by Color" and choose "No Fill." This way all the highlighted cells will be cleared out.

Result_6:

	A	B	C	D
1	County	State	Ethnicity	Population
2	Strafford County	Louisiana	Black African	440,000
3	Sherburne County	Kentucky	Black African	3,000,000
4	Marion County	North Dakota	White	5,200,000
5	Dickinson County	Iowa	Chinese	419,000
6	York County	Minnesota	Black African	4,000,000
7	Wallowa County	Oregon	Chinese	2,200,000
8	Iberia Parish	Indiana	Black African	15,950,000
9	Gillespie County	Oklahoma	Other Asian	850,000
10	Cavalier County	New York	Indian	420,000
11	Judith Basin County	Michigan	Black Caribbean	2,300,000
12	Cape Girardeau County	Iowa	Other Black	8,166,666
13	Randolph County	Illinois	Black African	417,000
14	Beauregard Parish	Illinois	Black African	3,000,000
15	Wharton County	South Carolina	Chinese	416,000
16	Pawnee County	Minnesota	Black African	2,000,000
17	Hall County	Texas	Black African	2,300,000
18	Crittenden County	Colorado	Chinese	440,000
19	Giles County	Ohio	Mixed	2,000,000
20	Fulton County	Alabama	Black African	7,750,000
21	Montgomery County	Idaho	Black African	2,000,000
22	Jackson County	Michigan	Chinese	19,000,000
23	Lassen County	Alabama	Other Black	425,000
24	Halifax County	Pennsylvania	Chinese	420,000

d) Data cleaning - Duplicate Values

To remove duplicate values, first click on any cell in the table and go to the "Table Design" tab, select "Remove Duplicates" and click "OK" to eliminate any duplicate values.

Result_7:

	A	B	C	D
1	County	State	Ethnicity	Population
2	Strafford County	Louisiana	Black African	440,000
3	Sherburne County	Kentucky	Black African	3,000,000
4	Marion County	North Dakota	White	5,200,000
5	Dickinson County	Iowa	Chinese	419,000
6	York County	Minnesota	Black African	4,000,000
7	Wallowa County	Oregon	Chinese	2,200,000
8	Iberia Parish	Indiana	Black African	15,950,000
9	Gillespie County	Oklahoma	Other Asian	850,000
10	Cavalier County	New York	Indian	420,000
11	Judith Basin County	Michigan	Black Caribbean	2,300,000
12	Cape Girardeau County	Iowa	Other Black	8,166,666
13	Randolph County	Illinois	Black African	417,000
14	Beauregard Parish	Illinois	Black African	3,000,000
15	Wharton County	South Carolina	Chinese	416,000
16	Pawnee County	Minnesota	Black African	2,000,000
17	Hall County	Texas	Black African	2,300,000
18	Crittenden County	Colorado	Chinese	440,000
19	Giles County	Ohio	Mixed	2,000,000
20	Fulton County	Alabama	Black African	7,750,000
21	Montgomery County	Idaho	Black African	2,000,000
22	Jackson County	Michigan	Chinese	19,000,000
23	Lassen County	Alabama	Other Black	425,000
24	Halifax County	Pennsylvania	Chinese	420,000

The above image shows that there are no duplicate values. So, with missing and incorrect entries removed, the dataset is now ready for analysis.

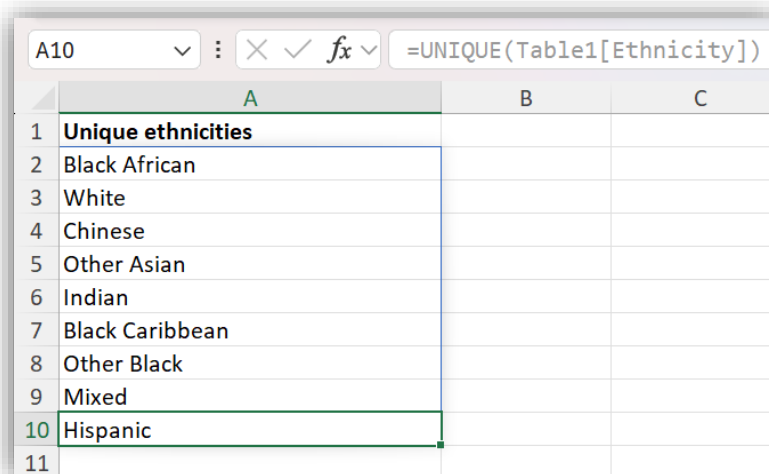
Standard statistics before and after data cleaning

Before data cleaning	After data cleaning																																																								
<table> <tr> <th colspan="2">Population</th></tr> <tr> <td>Mean</td><td>2902496.298</td></tr> <tr> <td>Standard Error</td><td>160524.231</td></tr> <tr> <td>Median</td><td>900000</td></tr> <tr> <td>Mode</td><td>414000</td></tr> <tr> <td>Standard Deviation</td><td>4660734.729</td></tr> <tr> <td>Sample Variance</td><td>2.17224E+13</td></tr> <tr> <td>Kurtosis</td><td>6.624052499</td></tr> <tr> <td>Skewness</td><td>1.722659854</td></tr> <tr> <td>Range</td><td>55125000</td></tr> <tr> <td>Minimum</td><td>-23125000</td></tr> <tr> <td>Maximum</td><td>32000000</td></tr> <tr> <td>Sum</td><td>2446804379</td></tr> <tr> <td>Count</td><td>843</td></tr> </table>	Population		Mean	2902496.298	Standard Error	160524.231	Median	900000	Mode	414000	Standard Deviation	4660734.729	Sample Variance	2.17224E+13	Kurtosis	6.624052499	Skewness	1.722659854	Range	55125000	Minimum	-23125000	Maximum	32000000	Sum	2446804379	Count	843	<table> <tr> <th colspan="2">Population</th></tr> <tr> <td>Mean</td><td>3288248.645</td></tr> <tr> <td>Standard Error</td><td>162463.4923</td></tr> <tr> <td>Median</td><td>1100000</td></tr> <tr> <td>Mode</td><td>414000</td></tr> <tr> <td>Standard Deviation</td><td>4522792.214</td></tr> <tr> <td>Sample Variance</td><td>2.04556E+13</td></tr> <tr> <td>Kurtosis</td><td>5.871119686</td></tr> <tr> <td>Skewness</td><td>2.272554731</td></tr> <tr> <td>Range</td><td>31586000</td></tr> <tr> <td>Minimum</td><td>414000</td></tr> <tr> <td>Maximum</td><td>32000000</td></tr> <tr> <td>Sum</td><td>2548392700</td></tr> <tr> <td>Count</td><td>775</td></tr> </table>	Population		Mean	3288248.645	Standard Error	162463.4923	Median	1100000	Mode	414000	Standard Deviation	4522792.214	Sample Variance	2.04556E+13	Kurtosis	5.871119686	Skewness	2.272554731	Range	31586000	Minimum	414000	Maximum	32000000	Sum	2548392700	Count	775
Population																																																									
Mean	2902496.298																																																								
Standard Error	160524.231																																																								
Median	900000																																																								
Mode	414000																																																								
Standard Deviation	4660734.729																																																								
Sample Variance	2.17224E+13																																																								
Kurtosis	6.624052499																																																								
Skewness	1.722659854																																																								
Range	55125000																																																								
Minimum	-23125000																																																								
Maximum	32000000																																																								
Sum	2446804379																																																								
Count	843																																																								
Population																																																									
Mean	3288248.645																																																								
Standard Error	162463.4923																																																								
Median	1100000																																																								
Mode	414000																																																								
Standard Deviation	4522792.214																																																								
Sample Variance	2.04556E+13																																																								
Kurtosis	5.871119686																																																								
Skewness	2.272554731																																																								
Range	31586000																																																								
Minimum	414000																																																								
Maximum	32000000																																																								
Sum	2548392700																																																								
Count	775																																																								

2.2

In this question, we need to plot the histogram of the population for each ethnicity. Therefore, we must first identify the number of ethnic groups in the dataset. To identify the unique ethnicities in the dataset, open a new worksheet, type “=UNIQUE(Table1[Ethnicity])” in a blank cell and press Enter. This will provide a list of the unique ethnicities present in the dataset.

Result_8:

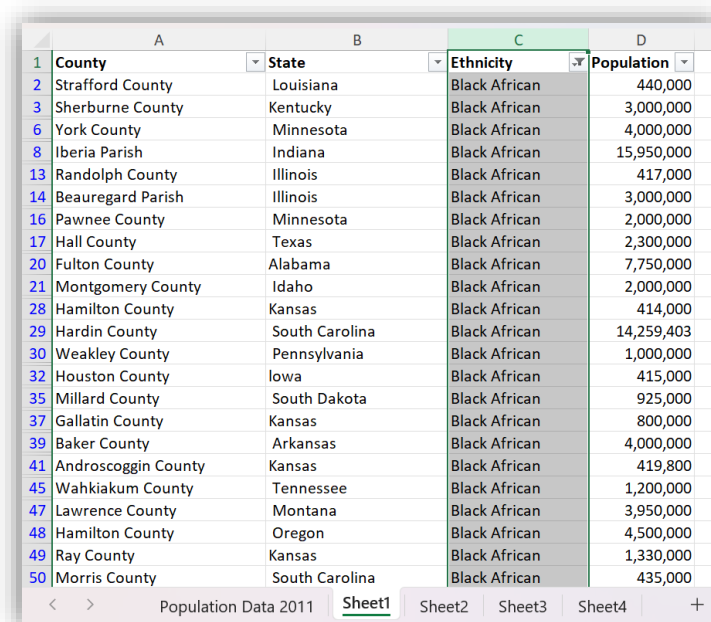


	A	B	C
1	Unique ethnicities		
2	Black African		
3	White		
4	Chinese		
5	Other Asian		
6	Indian		
7	Black Caribbean		
8	Other Black		
9	Mixed		
10	Hispanic		
11			

This image shows that there are 9 ethnicities in the dataset. So, we need to create 9 histograms. To do this, we first create separate tables for each ethnicity and then plot the corresponding histograms.

To create a new table containing Population of 'Black African', first click the white box on the right side of the 'Population' column in the table, give a tick only to 'Black African' checkbox, and click "OK". This will filter out other ethnicities and display only the "Black African" population.

Result_9:



	A	B	C	D
1	County	State	Ethnicity	Population
2	Stafford County	Louisiana	Black African	440,000
3	Sherburne County	Kentucky	Black African	3,000,000
6	York County	Minnesota	Black African	4,000,000
8	Iberia Parish	Indiana	Black African	15,950,000
13	Randolph County	Illinois	Black African	417,000
14	Beauregard Parish	Illinois	Black African	3,000,000
16	Pawnee County	Minnesota	Black African	2,000,000
17	Hall County	Texas	Black African	2,300,000
20	Fulton County	Alabama	Black African	7,750,000
21	Montgomery County	Idaho	Black African	2,000,000
28	Hamilton County	Kansas	Black African	414,000
29	Hardin County	South Carolina	Black African	14,259,403
30	Weakley County	Pennsylvania	Black African	1,000,000
32	Houston County	Iowa	Black African	415,000
35	Millard County	South Dakota	Black African	925,000
37	Gallatin County	Kansas	Black African	800,000
39	Baker County	Arkansas	Black African	4,000,000
41	Androscoggin County	Kansas	Black African	419,800
45	Wahkiakum County	Tennessee	Black African	1,200,000
47	Lawrence County	Montana	Black African	3,950,000
48	Hamilton County	Oregon	Black African	4,500,000
49	Ray County	Kansas	Black African	1,330,000
50	Morris County	South Carolina	Black African	435,000

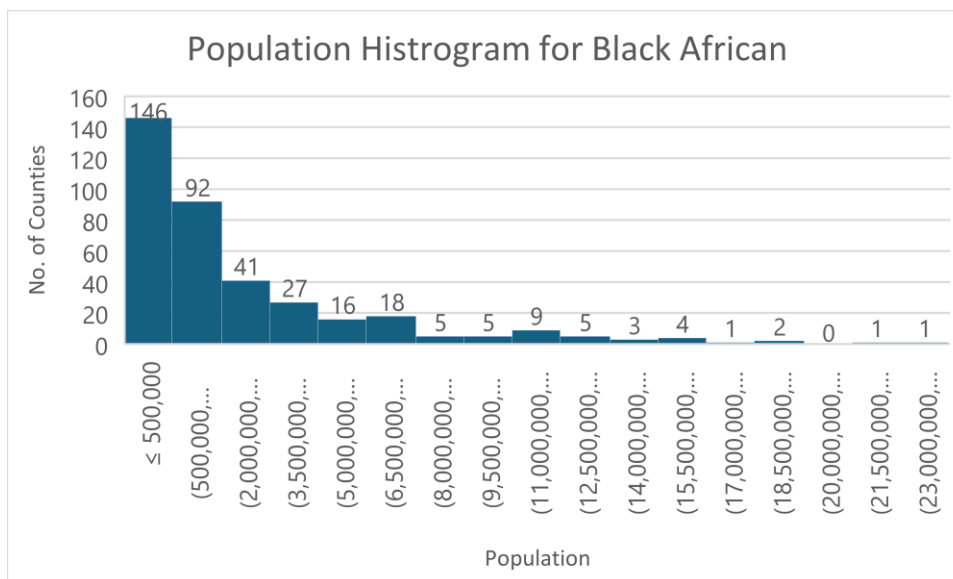
Then copy the entire table (A1 – D776) and paste it in the new worksheet.

Result_10:

	A	B	C	D
12	County	State	Ethnicity	Population
13	Strafford County	Louisiana	Black African	440,000
14	Sherburne County	Kentucky	Black African	3,000,000
15	York County	Minnesota	Black African	4,000,000
16	Iberia Parish	Indiana	Black African	15,950,000
17	Randolph County	Illinois	Black African	417,000
18	Beauregard Parish	Illinois	Black African	3,000,000
19	Pawnee County	Minnesota	Black African	2,000,000
20	Hall County	Texas	Black African	2,300,000
21	Fulton County	Alabama	Black African	7,750,000
22	Montgomery County	Idaho	Black African	2,000,000
23	Hamilton County	Kansas	Black African	414,000
24	Hardin County	South Carolina	Black African	14,259,403
25	Weakley County	Pennsylvania	Black African	1,000,000
26	Houston County	Iowa	Black African	415,000
27	Millard County	South Dakota	Black African	925,000
28	Gallatin County	Kansas	Black African	800,000
29	Baker County	Arkansas	Black African	4,000,000
30	Androscoggin County	Kansas	Black African	419,800
31	Wahkiakum County	Tennessee	Black African	1,200,000
32	Lawrence County	Montana	Black African	3,950,000
33	Hamilton County	Oregon	Black African	4,500,000
34	Ray County	Kansas	Black African	1,330,000
35	Morris County	South Carolina	Black African	435,000

Next select the population column of this copied table (D12:D388), go to "Insert", and choose "Histogram". Click on the graph and then click on the plus sign on the right side of the graph. Tick "Data Labels" and "Axis Titles" checkboxes and input the appropriate axis titles. Click above the horizontal axis title, right-click and select "Format Axis". Set the Bin Width to 1,500,000 and the Underflow bin to 500,000. Finally, change the chart title to identify the histogram.

Result_11:

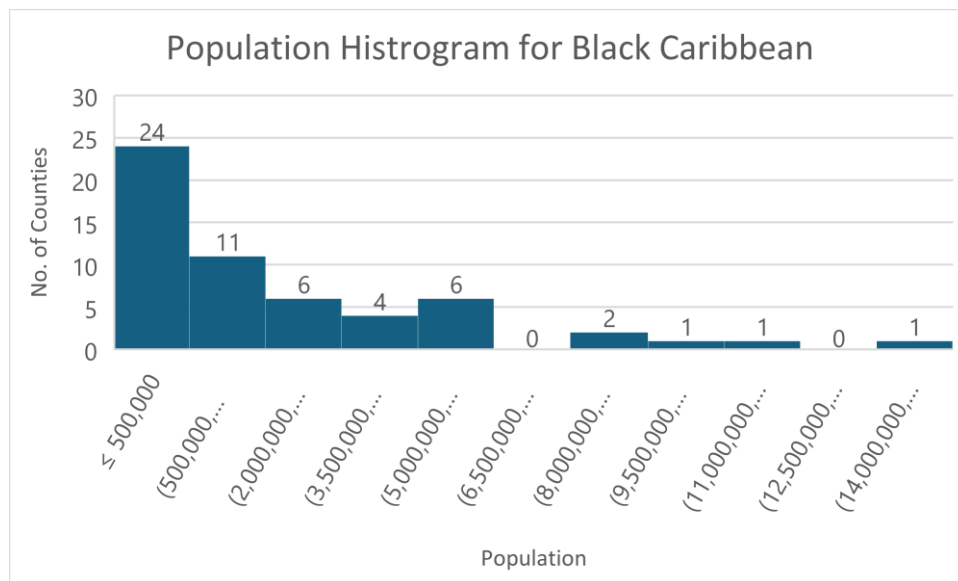


This graph shows the number of counties within each population range for the 'Black African' ethnicity. According to the graph most counties have 'Black African' below 500,000. That

is 146 counties. The highest population range is between 23,000,000 and 24,500,000. That is only 1 county.

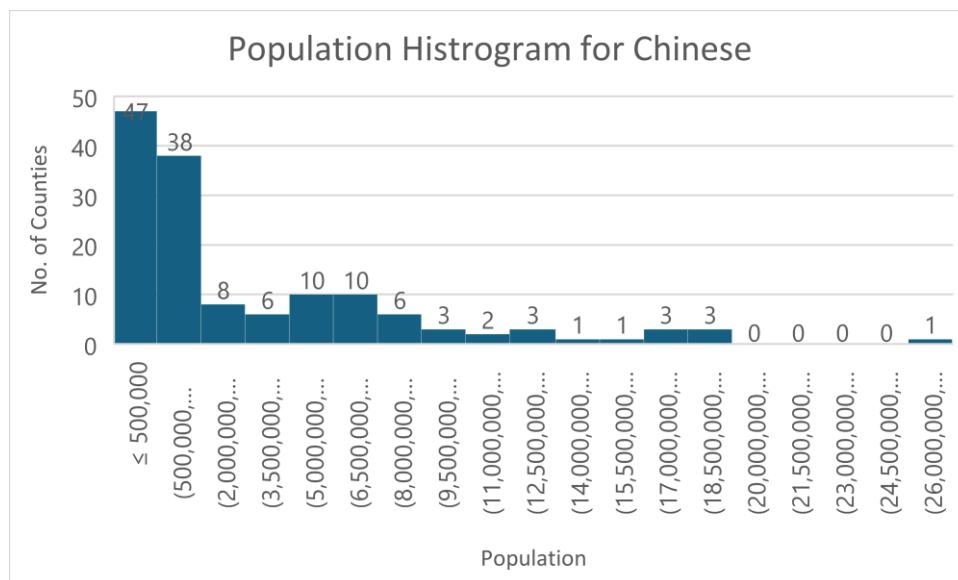
In this way we have created a histogram for each ethnicity.

Result_12:



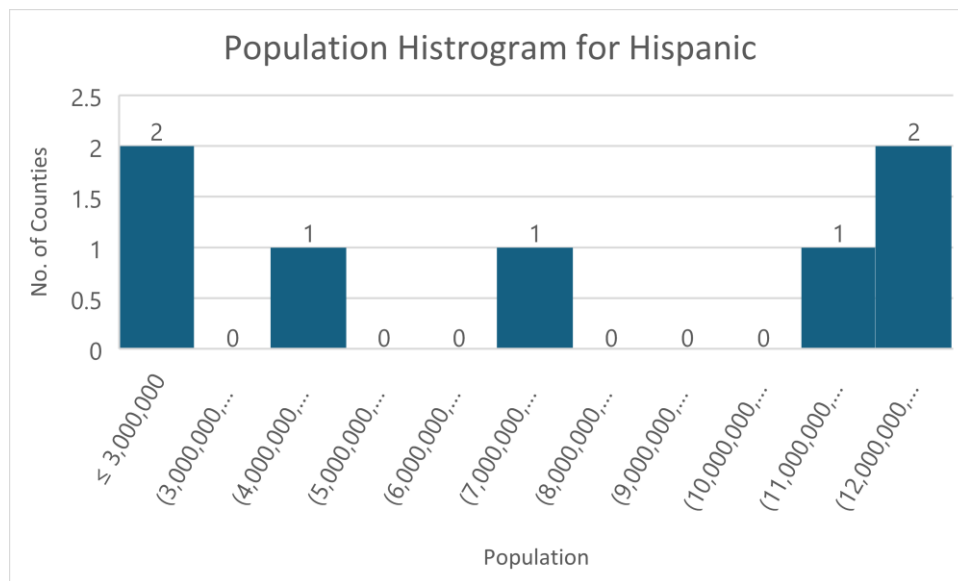
This graph shows the number of counties within each population range for the 'Black Caribbean' ethnicity. According to the graph most counties have 'Black Caribbean' population below 500,000. That is 24 counties. The highest population range is between 14,000,000 and 15,500,000. That is only 1 county. For this chart the bin width is 1,500,000 and the underflow bin is set to 500,000.

Result_13:



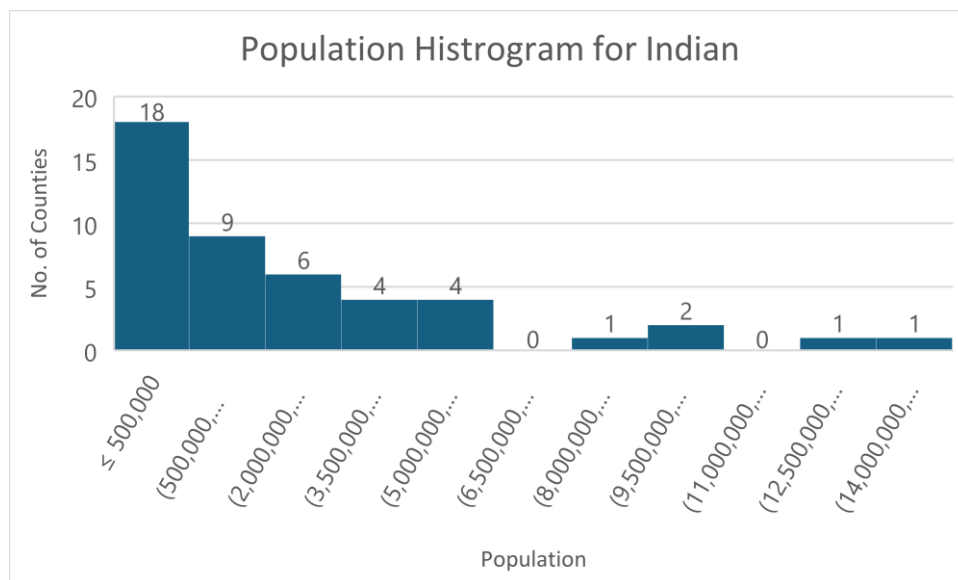
This graph shows the number of counties within each population range for the 'Chinese' ethnicity. According to the graph most counties have 'Chinese' population below 500,000. That is 47 counties. The highest population range is between 26,000,000 and 27,500,000. That is only 1 county. For this chart the bin width is 1,500,000 and the underflow bin is set to 500,000.

Result_14:



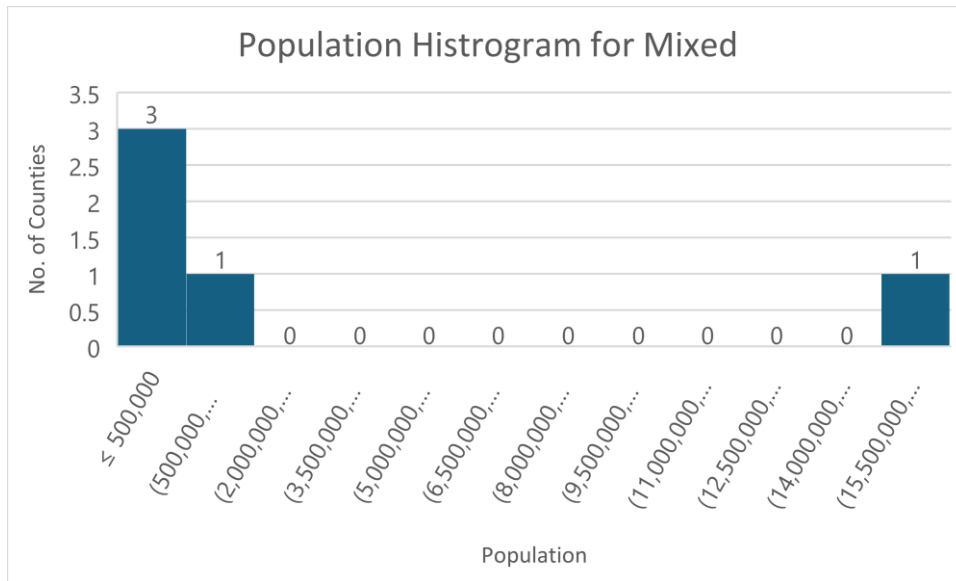
This graph shows the number of counties within each population range for the 'Hispanic' ethnicity. According to the graph, only 7 counties have a 'Hispanic' ethnic population. Of these, 2 counties have a population of less than 3,000,000 and 2 states have a population of more than 12,000,000. For this chart the bin width is 1,000,000 and the underflow bin is set to 3,000,000.

Result_15:



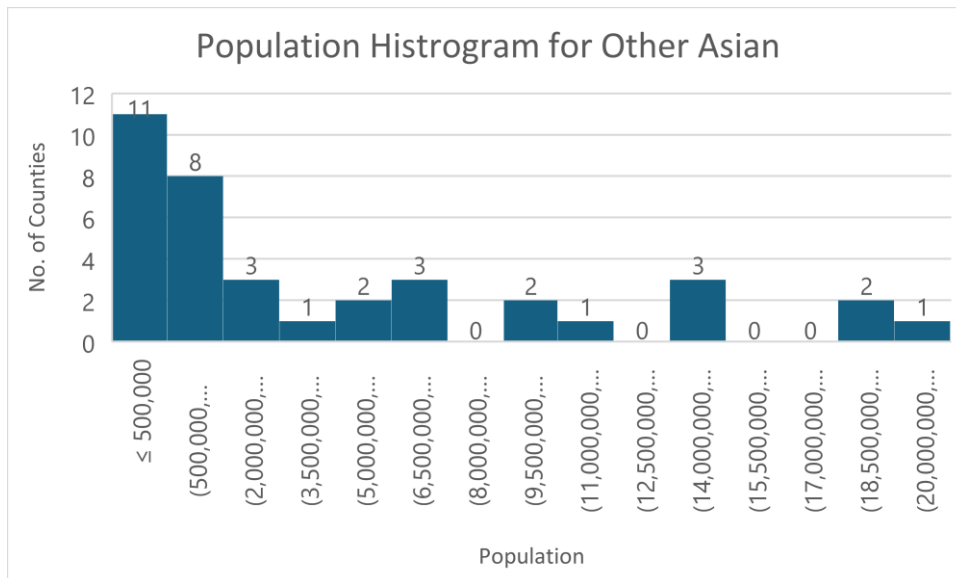
This graph shows the number of counties within each population range for the 'Indian' ethnicity. According to the graph most counties have 'Indian' population below 500,000. That is 18 counties. The highest population range is between 14,000,000 and 15,500,000. That is only 1 county. For this chart the bin width is 1,500,000 and the underflow bin is set to 500,000.

Result_16:



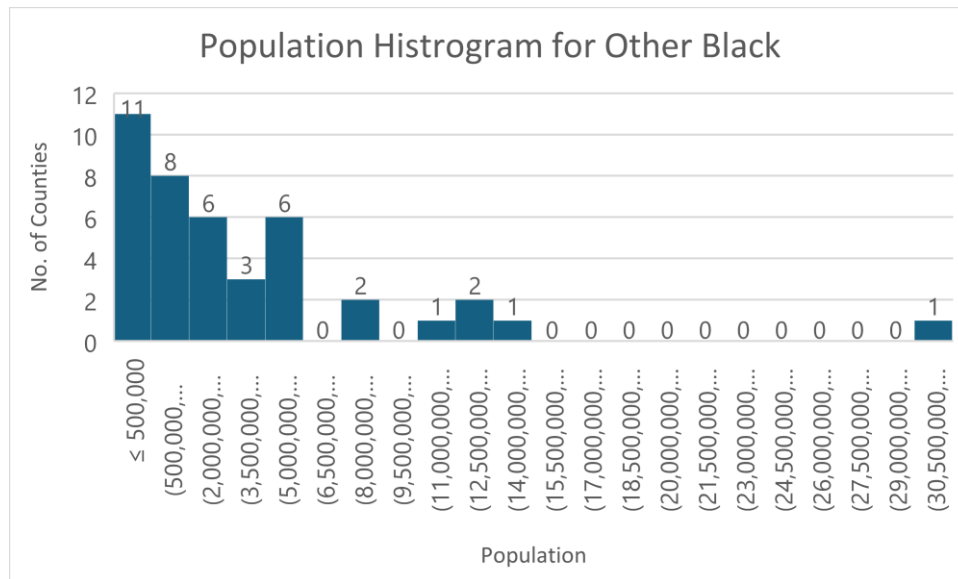
This graph shows the number of counties within each population range for the 'Mixed' ethnicity. According to the graph most counties have 'Mixed' population below 500,000. That is 3 counties. The highest population range is between 15,500,000 and 17,000,000. That is only 1 county. According to the graph, only 5 counties have a 'Mixed' ethnic population. For this chart the bin width is 1,500,000 and the underflow bin is set to 500,000.

Result_17:



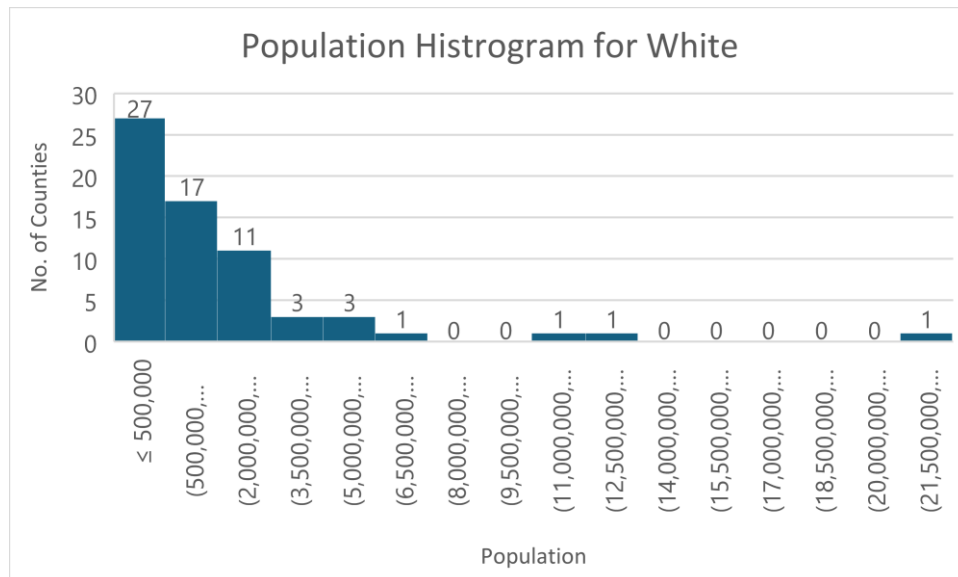
This graph shows the number of counties within each population range for the 'Other Asian' ethnicity. According to the graph most counties have 'Other Asian' population below 500,000. That is 11 counties. The highest population range is between 20,000,000 and 21,500,000. That is only 1 county. For this chart the bin width is 1,500,000 and the underflow bin is set to 500,000.

Result_18:



This graph shows the number of counties within each population range for the 'Other Black' ethnicity. According to the graph most counties have 'Other Black' population below 500,000. That is 11 counties. The highest population range is between 30,500,000 and 32,000,000. That is only 1 county. For this chart the bin width is 1,500,000 and the underflow bin is set to 500,000.

Result_19:



This graph shows the number of counties within each population range for the 'White' ethnicity. According to the graph most counties have 'White' population below 500,000. That is 27 counties. The highest population range is between 21,500,000 and 23,000,000. That is only 1 county. For this chart the bin width is 1,500,000 and the underflow bin is set to 500,000.

2.3

Descriptive statistics for each position

To find descriptive statistics for each ethnicity and as a whole, go to the Data tab and select Data Analysis. Then choose Descriptive Statistics and click OK. Set the Input Range to include the relevant data range and tick the Labels in First Row checkbox. Finally, select the Summary Statistics option and click OK to generate the descriptive statistics for each position.

1. Black African Ethnicity
Input Range: Sheet2!\$D\$12:\$D\$388

Result_20:

1. Black african	
Population	
Mean	2957133.721
Standard Error	210301.3965
Median	1087500
Mode	414000
Standard Deviation	4077895.375
Sample Variance	1.66292E+13
Kurtosis	5.599837815
Skewness	2.295449572
Range	23871714
Minimum	414000
Maximum	24285714
Sum	1111882279
Count	376

2. Black Caribbean Ethnicity
Input Range: Sheet2!\$D\$392:\$D\$448

Result_21:

2. Black Caribbean	
Population	
Mean	2603493.429
Standard Error	443790.7038
Median	825000
Mode	414000
Standard Deviation	3321025.53
Sample Variance	1.10292E+13
Kurtosis	3.568941937
Skewness	1.892765665
Range	14871714
Minimum	414000
Maximum	15285714
Sum	145795632
Count	56

3. Chinese Ethnicity
Input Range: Sheet2!\$D\$452:\$D\$594

Result_22:

3. Chinese	
Population	
Mean	3917210.887
Standard Error	435730.2001
Median	1137500
Mode	414000
Standard Deviation	5192324.589
Sample Variance	2.69602E+13
Kurtosis	3.265921285
Skewness	1.862090213
Range	25773500
Minimum	414000
Maximum	26187500
Sum	556243946
Count	142

4. Hispanic Ethnicity
Input Range: Sheet2!\$D\$598:\$D\$605

Result_23:

4. Hispanic	
Population	
Mean	7768779.286
Standard Error	1798475.718
Median	7611455
Mode	#N/A
Standard Deviation	4758319.488
Sample Variance	2.26416E+13
Kurtosis	-2.365783761
Skewness	-0.037240619
Range	10980000
Minimum	2020000
Maximum	13000000
Sum	54381455
Count	7

5. Indian Ethnicity
Input Range: Sheet2!\$D\$609:\$D\$655

Result_24:

5. Indian	
Population	
Mean	2870720.326
Standard Error	531054.9351
Median	1175000
Mode	414000
Standard Deviation	3601789.809
Sample Variance	1.29729E+13
Kurtosis	2.981365022
Skewness	1.854310665
Range	14315364
Minimum	414000
Maximum	14729364
Sum	132053135
Count	46

6. Mixed Ethnicity
Input Range: Sheet2!\$D\$659:\$D\$664

Result_25:

6.Mixed	
Population	
Mean	3887814.8
Standard Error	3086964.695
Median	428600
Mode	#N/A
Standard Deviation	6902662.901
Sample Variance	4.76468E+13
Kurtosis	4.795666727
Skewness	2.183300223
Range	15760974
Minimum	414000
Maximum	16174974
Sum	19439074
Count	5

7. Other Asian Ethnicity
Input Range: Sheet2!\$D\$668:\$D\$705

Result_26:

7. Other Asian	
Population	
Mean	5397852.595
Standard Error	1042488.32
Median	2000000
Mode	414000
Standard Deviation	6341208.893
Sample Variance	4.02109E+13
Kurtosis	0.343859099
Skewness	1.233621911
Range	19861000
Minimum	414000
Maximum	20275000
Sum	199720546
Count	37

8. Other Black Ethnicity
Input Range: Sheet2!\$D\$709:\$D\$750

Result_27:

8. Other Black	
Population	
Mean	4404601.585
Standard Error	937364.2976
Median	2287500
Mode	414000
Standard Deviation	6002060.053
Sample Variance	3.60247E+13
Kurtosis	10.54860408
Skewness	2.852149057
Range	31586000
Minimum	414000
Maximum	32000000
Sum	180588665
Count	41

9. White Ethnicity
Input Range: Sheet2!\$D\$754:\$D\$819

Result_28:

9. White	
Population	
Mean	2281353.354
Standard Error	447818.604
Median	850000
Mode	3000000
Standard Deviation	3610429.01
Sample Variance	1.30352E+13
Kurtosis	17.88062501
Skewness	3.808787173
Range	22586000
Minimum	414000
Maximum	23000000
Sum	148287968
Count	65

10. All Ethnicities
Input Range: Sheet1!\$D\$1:\$D\$776

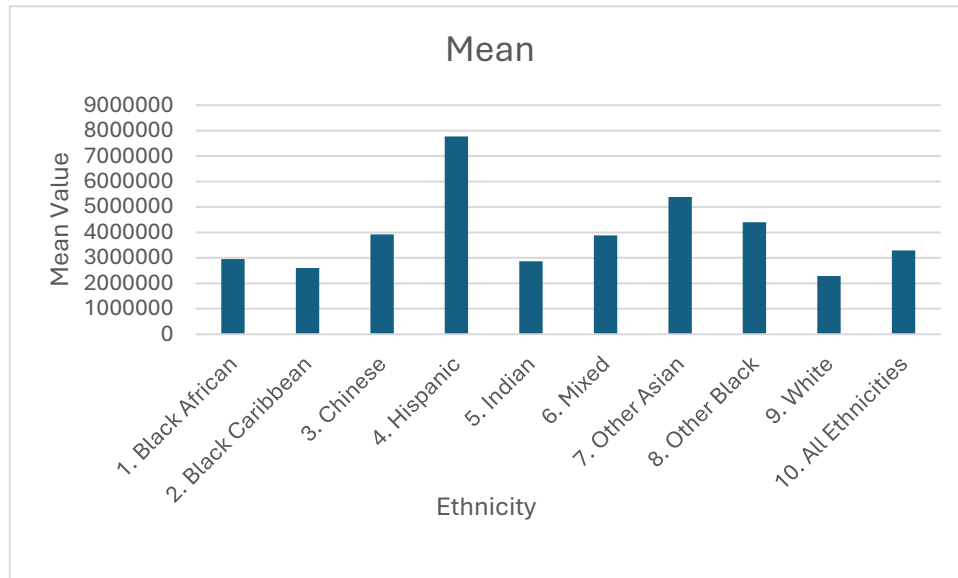
Result_29:

10. All Ethnicities	
Population	
Mean	3288248.645
Standard Error	162463.4923
Median	1100000
Mode	414000
Standard Deviation	4522792.214
Sample Variance	2.04556E+13
Kurtosis	5.871119686
Skewness	2.272554731
Range	31586000
Minimum	414000
Maximum	32000000
Sum	2548392700
Count	775

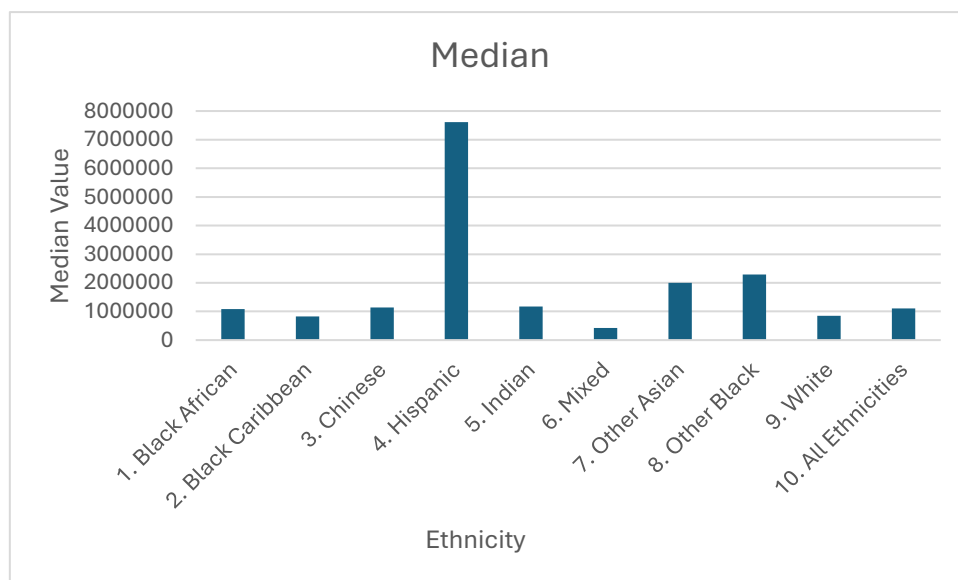
Comparative table

Result_30:

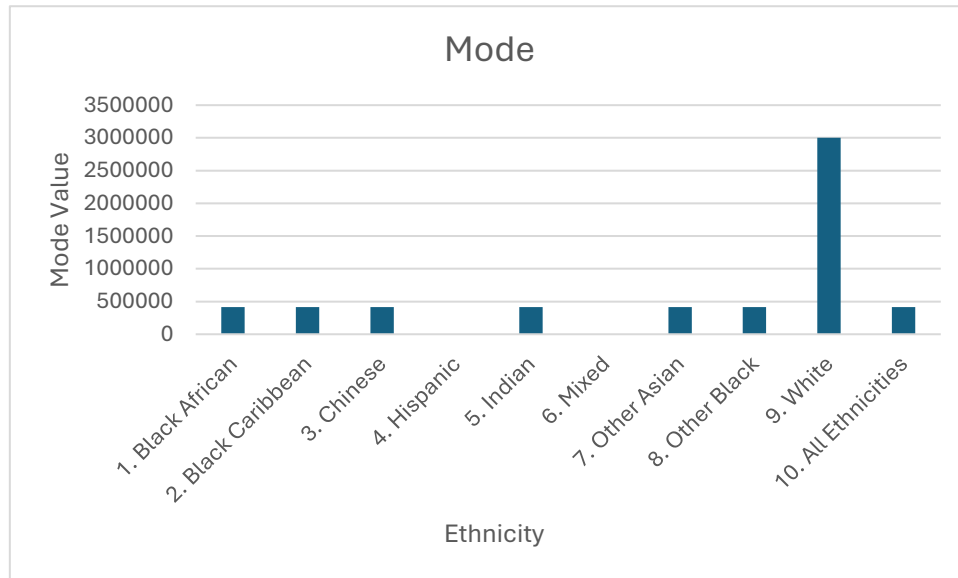
		1. Black african	2. Black Caribbean	3. Chinese	4. Hispanic	5. Indian	6.Mixed	7. Other Asian	8. Other Black	9. White	10. All Ethnicities
1	Mean	2957133.721	2603493.429	3917210.887	7768779.286	2870720.326	3887814.8	5397852.595	4404601.585	2281353.354	3288248.645
2	Median	1087500	825000	1137500	7611455	1175000	428600	2000000	2287500	850000	1100000
3	Mode	414000	414000	414000	#N/A	414000	#N/A	414000	414000	3000000	414000
4	Standard Deviation	4077895.375	3321025.53	5192324.589	4758319.488	3601789.809	6902662.901	6341208.893	6002060.053	3610429.01	4522792.214
5	Sample Variance	1.66292E+13	1.10292E+13	2.69602E+13	2.26416E+13	1.29729E+13	4.76468E+13	4.02109E+13	3.60247E+13	1.30352E+13	2.04556E+13
6	Range	23871714	14871714	25773500	10980000	14315364	15760974	19861000	31586000	22586000	31586000
7	Minimum	414000	414000	414000	2020000	414000	414000	414000	414000	414000	414000
8	Maximum	24285714	15285714	26187500	13000000	14729364	16174974	20275000	32000000	23000000	32000000
9	Sum	1111882279	145795632	556243946	54381455	132053135	19439074	199720546	180588665	148287968	2548392700
10	Count	376	56	142	7	46	5	37	41	65	775



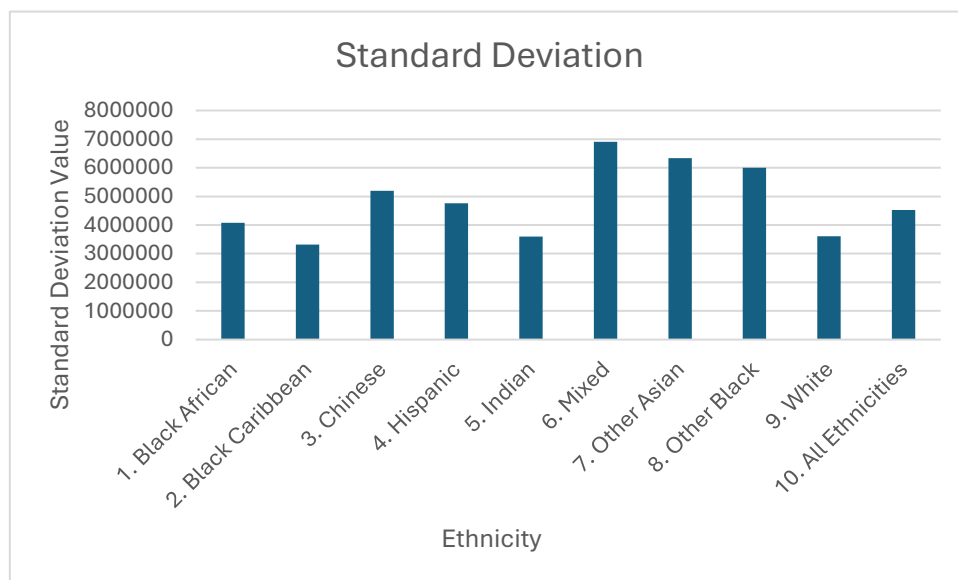
The bar graph of mean populations by ethnicity reveals that Hispanics have the highest mean, followed by 'Other Asian' and 'Other Black' ethnicities. The 'Chinese' and 'Mixed' ethnicities have similar means near 4,000,000. 'Indian' and 'Black African' populations are just under 3,000,000. 'White' ethnicity has the lowest mean value. The overall mean population is slightly higher than 3,000,000 indicating most ethnicities are around or above this average, with 'Hispanics' significantly higher.



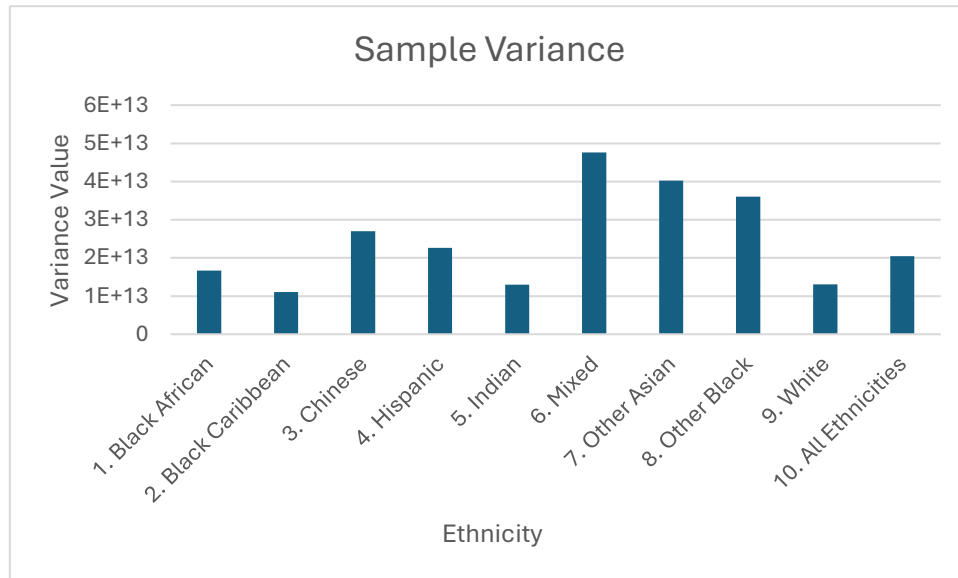
The bar graph of median populations by ethnicity highlights significant differences. 'Hispanic' has the highest median at over 7,000,000. 'Other Black' and 'Other Asian' follow with medians near 2,000,000. 'Indian' and 'Chinese' populations have medians slightly above the overall median of 1,100,000. 'Black African' and 'White' ethnicities are slightly below the overall median. 'Mixed' ethnicities have the lowest medians. This distribution emphasizes the notably higher 'Hispanic' median compared to other ethnicities.



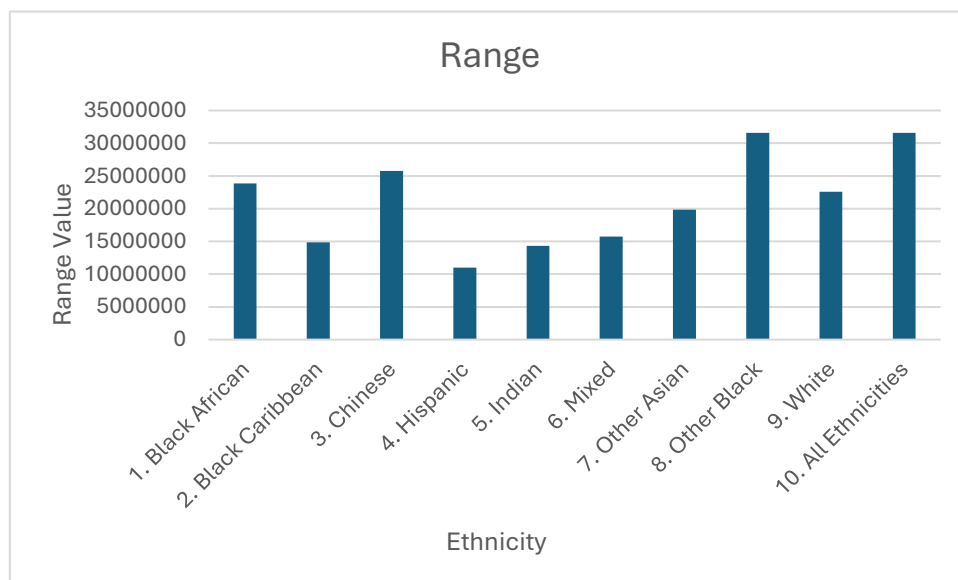
The bar graph of population modes shows most ethnicities; 'Black African', 'Black Caribbean', 'Chinese', 'Indian', 'Other Asian' and 'Other Black' sharing a mode of 414,000. The 'White' ethnicity stands out with a mode of 3,000,000. The 'Hispanic' and 'Mixed' ethnicities have no mode. Overall, the mode for all ethnicities is 414,000, indicating uniformity across several groups, with the 'White' ethnicity being notably higher.



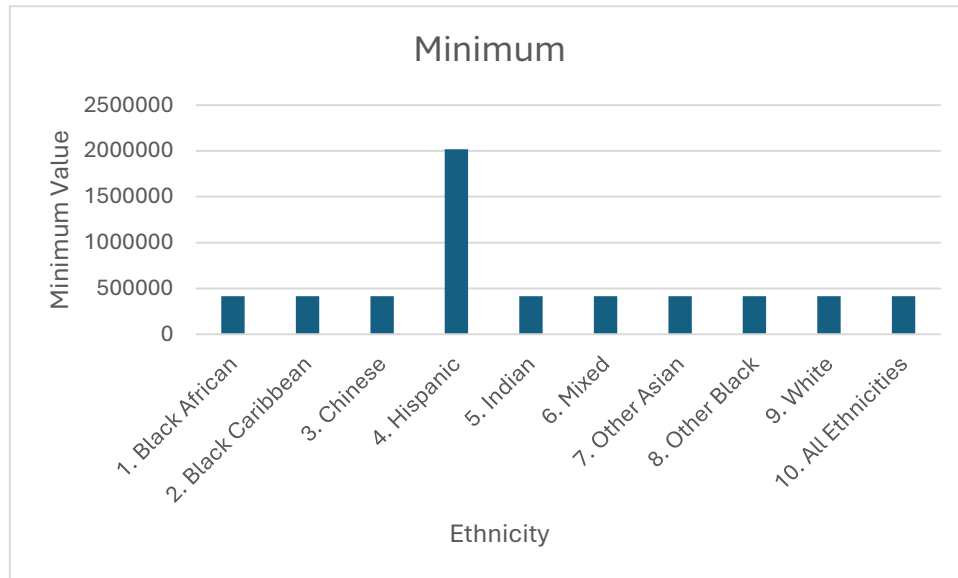
The bar graph shows the standard deviations in population sizes with the 'Mixed' ethnicities having the highest at slightly lower than 7,000,000. 'Other Asian' and 'Other Black' follow with slightly higher than 6,000,000. The 'Black Caribbean' ethnicity has the lowest at slightly higher than 3,000,000. Overall, the standard deviation across all ethnicities is between 4,000,000 and 5,000,000. This highlights the diverse population sizes within the ethnicities.



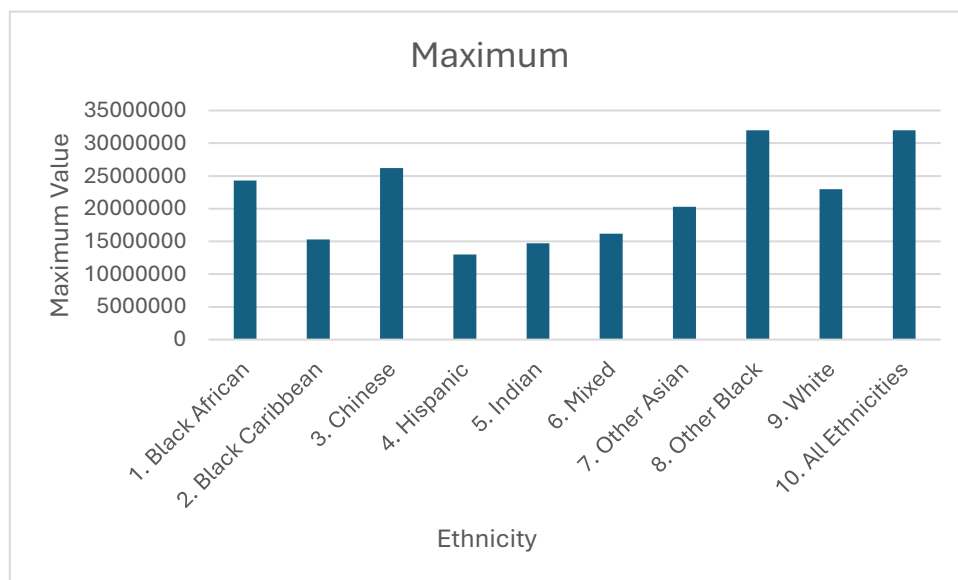
The bar graph of sample variance in population sizes shows the 'Mixed' ethnicity has the highest variance value. 'Other Asian' and 'Other Black' follow closely behind it. 'Black Caribbean' has the lowest variance value. Overall, the variance across all ethnicities is slightly higher than the 2E+13 (20,000,000,000,000).



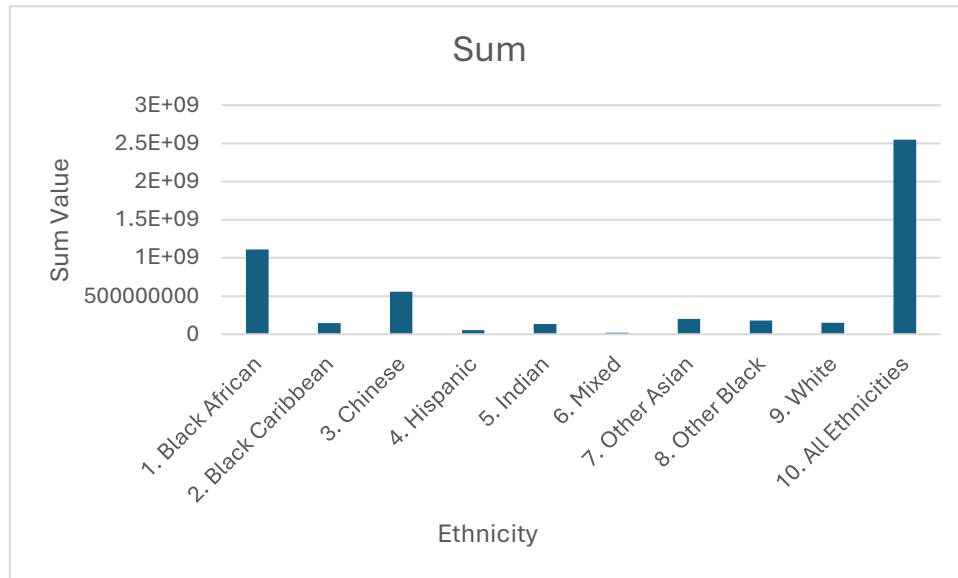
The bar graph reveals that 'Other Black' has the highest population at above 30,000,000. 'Chinese' follows with slightly above 25,000,000, while 'Black African' and 'White' slightly lower than 25,000,000. The smallest population range belong to 'Hispanic' with being slightly higher than 10,000,000.



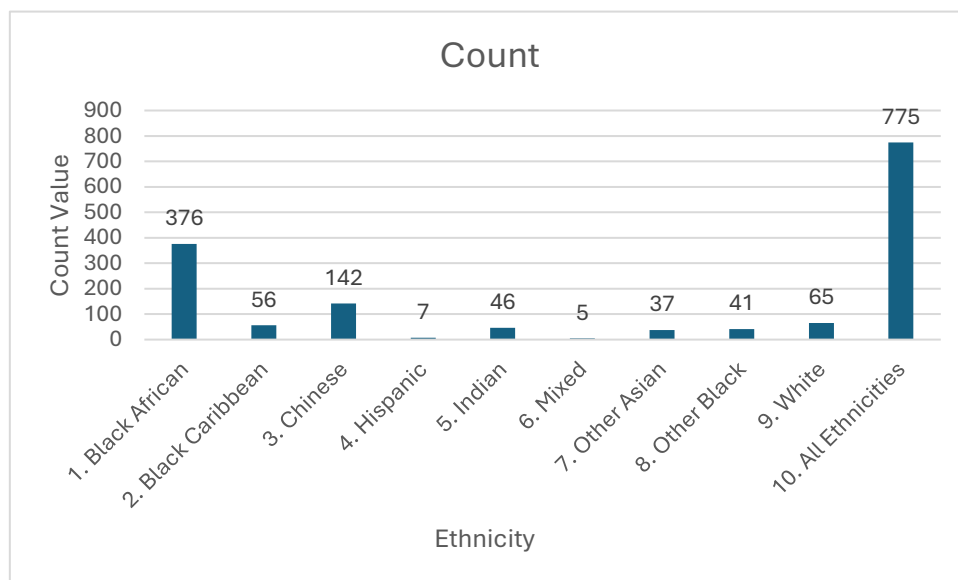
In the bar graph displaying the minimum population figures for each ethnicity, it is evident that most ethnic groups have the same minimum population value of 414,000. The exception is the 'Hispanic' ethnicity, which has a significantly higher minimum population of 2,020,000.



The bar graph illustrates the maximum population figures for each ethnicity. The 'Other Black' has the highest maximum population reaching above 30,000,000. Following closely are the 'Chinese' and 'Black African' ethnicities with populations slightly higher and slightly lower than 25,000,000 respectively. Conversely, "Hispanic" and "Indian" ethnicities have lowest maximum values below the 15,000,000. This distribution underscores the diverse range of population sizes among different ethnic groups.



The bar graph shows the sum of populations for each ethnicity. The 'Black African' category has the highest total population sum above 1,000,000,000 and 'Mixed' ethnicity has the smallest total sum when compared to other ethnicities. The second highest total population sum belong to 'Chinese' with slightly above 500,000,000. The total population in all ethnicities exceeds 2,500,000,000.



The bar graph illustrates the count of the population for each ethnicity. The "Black African" category has the highest count with 376 records, indicating the most extensive data set among the ethnic groups. In contrast, the "Mixed" ethnicity has the fewest records, with only 5. Other ethnicities show varying counts, with "Chinese" at 142, "Other Asian" at 37, and "White" at 65. The total number of records across all ethnicities is 775, illustrating the distribution of data across the different ethnic categories.

2.4

In this question, we will create charts to illustrate both the number and proportion of counties per ethnicity. This is by using two types of graphs: one representing absolute frequencies and the other representing relative frequencies.

First, we will create a new table in a new work sheet. By using the formula “=UNIQUE(Sheet1!\$C\$2:\$C\$776)”, we will obtain the unique values from the 'Ethnicity' column in the sheet 1 and list them under the 'Ethnicity' column of the new work sheet. Next, we will add a column named 'Absolute Count' and use the formula “=COUNTIF(Sheet1!\$C\$2:\$C\$776,B3)” to count the number of counties per ethnicity. Then find total using “=SUM(C3:C11)”. Then, we will create another column named 'Relative Count' and use the formula “=C3/\$C\$12” to calculate the percentage value for each position and apply the 'Percent Style' from the 'Home' tab to them.

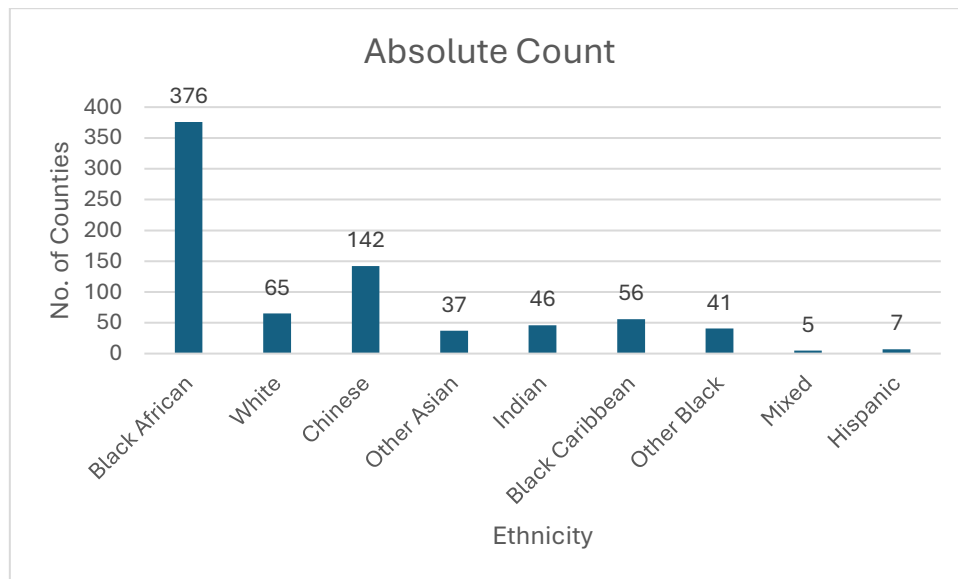
Result_31:

Ethnicity	Absolute Count	Relative Count
Black African	376	48.5%
White	65	8.4%
Chinese	142	18.3%
Other Asian	37	4.8%
Indian	46	5.9%
Black Caribbean	56	7.2%
Other Black	41	5.3%
Mixed	5	0.6%
Hispanic	7	0.9%
Total	775	100.0%

Next create a bar chart and pie chart for absolute count:

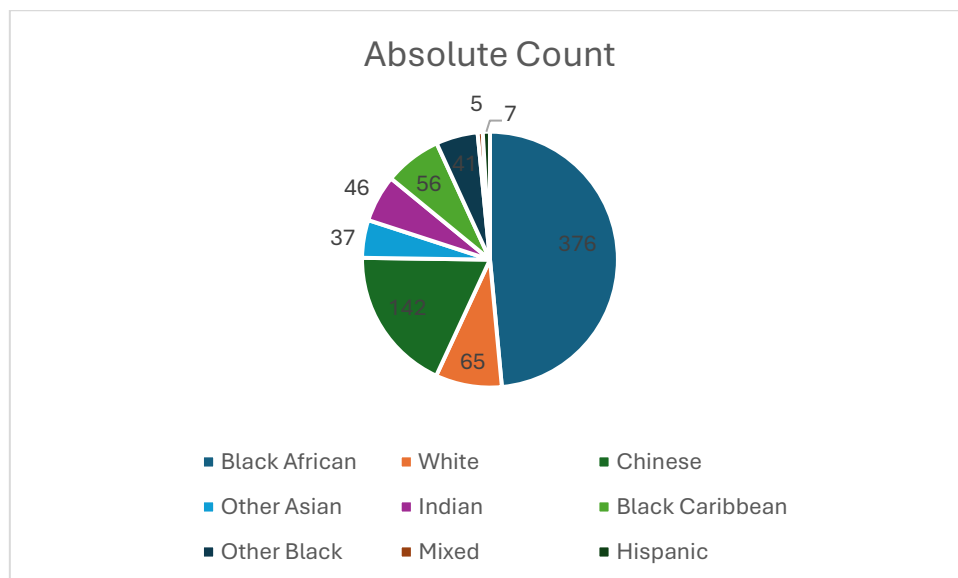
Select the entire table's first 2 columns except last row (B2:C11) and go to the Insert tab, then choose 'Insert Column or Bar Chart', and select '2-D Clustered Column' in the '2-D Column'. Click on the graph and then click on the plus icon on the right side of the graph. Then tick 'Data Labels' and 'Axis Titles' and input the appropriate titles for each axis.

Result_32:



Next, select table's first 2 columns except last row (B2:C11) again and go to the Insert tab, then choose 'Insert Pie or Doughnut Chart', and select 'Pie' of the '2-D Pie'.

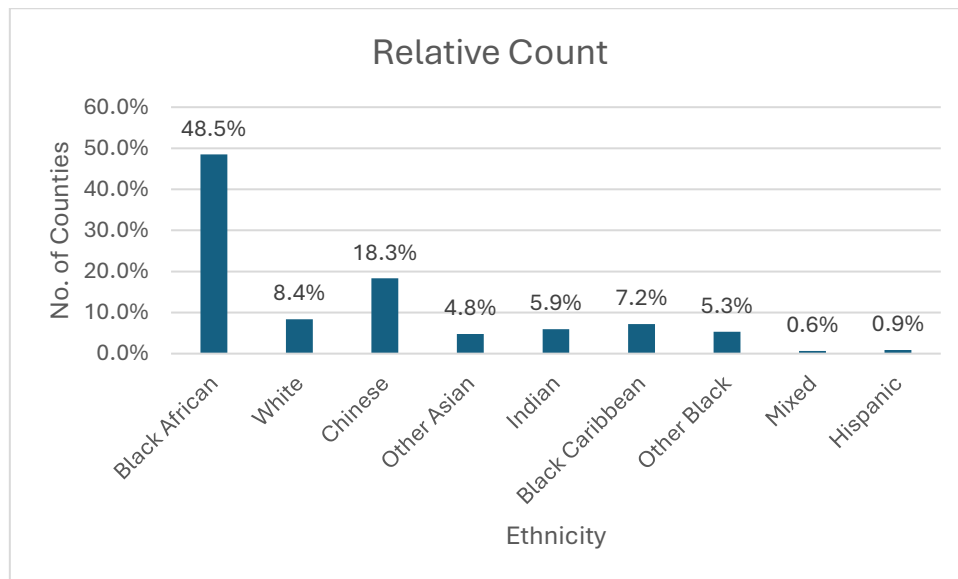
Result_33:



Now create a bar chart and pie chart for relative count:

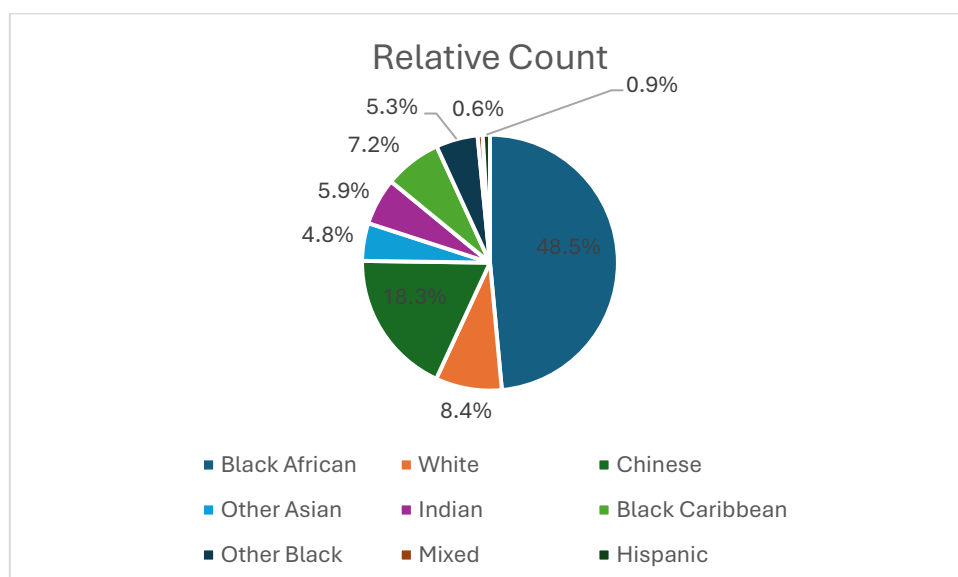
Select the entire table except middle column and last row (B2:B11, D2:D11) and go to the Insert tab, then choose 'Insert Column or Bar Chart', and select '2-D Clustered Column' in the '2-D Column'. Click on the graph and then click on the plus icon on the right side of the graph. Then tick 'Data Labels' and 'Axis Titles' and input the appropriate titles for each axis.

Result_34:



Next, Select the entire table except middle column and last row (B2:B11, D2:D11) again and go to the Insert tab, then choose 'Insert Pie or Doughnut Chart', and select 'Pie' of the '2-D Pie'.

Result_35:



Part 3 : Motor Trend Car Road Test Data Analysis

3.1

R: Motor Trend Car Road Tests ▾ Find in Topic

mtcars {datasets} R Documentation

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Usage

```
mtcars
```

Format

A data frame with 32 observations on 11 (numeric) variables.

- [, 1] `mpg` Miles/(US) gallon
- [, 2] `cyl` Number of cylinders
- [, 3] `disp` Displacement (cu.in.)
- [, 4] `hp` Gross horsepower
- [, 5] `drat` Rear axle ratio
- [, 6] `wt` Weight (1000 lbs)
- [, 7] `qsec` 1/4 mile time
- [, 8] `vs` Engine (0 = V-shaped, 1 = straight)
- [, 9] `am` Transmission (0 = automatic, 1 = manual)
- [, 10] `gear` Number of forward gears
- [, 11] `carb` Number of carburetors

Note

Henderson and Velleman (1981) comment in a footnote to Table 1: 'Hocking [original transcriber]'s noncrucial coding of the Mazda's rotary engine as a straight six-cylinder engine and the Porsche's flat engine as a V engine, as well as the inclusion of the diesel Mercedes 240D, have been retained to enable direct comparisons to be made with previous analyses.'

Source

Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, 37, 391–411.

Examples

[Run examples](#)

```
require(graphics)
pairs(mtcars, main = "mtcars data", gap = 1/4)
coplot(mpg ~ disp | as.factor(cyl), data = mtcars,
       panel = panel.smooth, rows = 1)
## possibly more meaningful, e.g., for summary() or bivariate plots:
mtcars2 <- within(mtcars, {
  vs <- factor(vs, labels = c("V", "S"))
  am <- factor(am, labels = c("automatic", "manual"))
  cyl <- ordered(cyl)
  gear <- ordered(gear)
  carb <- ordered(carb)
})
summary(mtcars2)
```

[Package *datasets* version 4.4.0 [Index](#)]

Code

```
1 #Question_3.1
2 data(mtcars) #Load mtcars dataset
3 ?mtcars #get description
4 colnames(mtcars) #indicate the variable names
5 str(mtcars) #display variable names,their class and size of vectors
6 head(mtcars,n=6) #showing first 6 lines
7 |
```

Output

```
Console Terminal Background Jobs
R 4.4.0 · D:/IIT/Course materials/2 Sem/Statistical Modelling and Analysis/Assesment/CW Refer/
> #Question_3.1
> data(mtcars) #Load mtcars dataset
> ?mtcars #get description
> colnames(mtcars) #indicate the variable names
[1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear"
[11] "carb"
> str(mtcars) #display variable names,their class and size of vectors
'data.frame': 32 obs. of 11 variables:
 $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num 160 160 108 258 360 ...
 $ hp : num 110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt : num 2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num 16.5 17 18.6 19.4 17 ...
 $ vs : num 0 0 1 1 0 1 0 1 1 1 ...
 $ am : num 1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num 4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num 4 4 1 1 2 1 4 2 2 4 ...
> head(mtcars,n=6) #showing first 6 lines
      mpg  cyl  disp  hp drat   wt  qsec vs  am gear carb
Mazda RX4    21.0   6  160  110 3.90 2.620 16.46 0   1    4    4
Mazda RX4 Wag 21.0   6  160  110 3.90 2.875 17.02 0   1    4    4
Datsun 710   22.8   4  108   93 3.85 2.320 18.61 1   1    4    1
Hornet 4 Drive 21.4   6  258  110 3.08 3.215 19.44 1   0    3    1
Hornet Sportabout 18.7   8  360  175 3.15 3.440 17.02 0   0    3    2
Valiant     18.1   6  225  105 2.76 3.460 20.22 1   0    3    1
> |
```

To get the information on 'mtcars' Dataset, first we need to load the dataset by using data (mtcars) function to the R. By executing '?mtcars' function, it provides detailed information about the dataset's background, structure and the variable. The 'colnames(mtcars)' function returns the names of the columns in the dataset, which are 'mpg', 'cyl', 'disp', 'hp', 'drat', 'wt', 'qsec', 'vs', 'am', 'gear' and 'carb'. The str(mtcars) command then displays the structure of the dataset, a data frame with 32 observations and 11 variables, where all are numerical variables. Finally, the head (mtcars, n=6) function gives the first six rows of the dataset. These commands help you understand what the 'mtcars' dataset looks like and what's inside it.

3.2

standard descriptive statistics

Central Statistics

	Code	Output
1. Mean 2. standard	<pre>#Mean mean(mtcars\$mpg) mean(mtcars\$cyl) mean(mtcars\$disp) mean(mtcars\$hp) mean(mtcars\$drat) mean(mtcars\$wt) mean(mtcars\$qsec) mean(mtcars\$vs) mean(mtcars\$am) mean(mtcars\$gear) mean(mtcars\$carb)</pre> <p>Here, 'mean(mtcars\$mpg)' calculates the mean of the 'mpg' column in the 'mtcars' dataset mean() function in RStudio.</p>	<pre>> #Mean > mean(mtcars\$mpg) [1] 20.09062 > mean(mtcars\$cyl) [1] 6.1875 > mean(mtcars\$disp) [1] 230.7219 > mean(mtcars\$hp) [1] 146.6875 > mean(mtcars\$drat) [1] 3.596563 > mean(mtcars\$wt) [1] 3.21725 > mean(mtcars\$qsec) [1] 17.84875 > mean(mtcars\$vs) [1] 0.4375 > mean(mtcars\$am) [1] 0.40625 > mean(mtcars\$gear) [1] 3.6875 > mean(mtcars\$carb) [1] 2.8125 ></pre>
2. Median	<pre>#Median median(mtcars\$mpg) median(mtcars\$cyl) median(mtcars\$disp) median(mtcars\$hp) median(mtcars\$drat) median(mtcars\$wt) median(mtcars\$qsec) median(mtcars\$vs) median(mtcars\$am) median(mtcars\$gear) median(mtcars\$carb)</pre> <p>Here, 'median(mtcars\$mpg)' calculates the median of the 'mpg' column in the 'mtcars' dataset median() function in RStudio.</p>	<pre>> #Median > median(mtcars\$mpg) [1] 19.2 > median(mtcars\$cyl) [1] 6 > median(mtcars\$disp) [1] 196.3 > median(mtcars\$hp) [1] 123 > median(mtcars\$drat) [1] 3.695 > median(mtcars\$wt) [1] 3.325 > median(mtcars\$qsec) [1] 17.71 > median(mtcars\$vs) [1] 0 > median(mtcars\$am) [1] 0 > median(mtcars\$gear) [1] 4 > median(mtcars\$carb) [1] 2</pre>

3. Mode

```
#Mode
#Assigned to variables in a table
A1 <- table(mtcars$mpg)
mode_mpg <- c(names(A1)[A1 == max(A1)])
mode_mpg
A2<-table(mtcars$cyl)
mode_cyl <- c(names(A2)[A2 == max(A2)])
mode_cyl
A3<-table(mtcars$disp)
mode_disp <- c(names(A3)[A3 == max(A3)])
mode_disp
A4<-table(mtcars$hp)
mode_hp <- c(names(A4)[A4 == max(A4)])
mode_hp
A5<-table(mtcars$drat)
mode_drat <- c(names(A5)[A5 == max(A5)])
mode_drat
A6<-table(mtcars$wt)
mode_wt <- c(names(A6)[A6 == max(A6)])
mode_wt
A7<-table(mtcars$qsec)
mode_qsec <- c(names(A7)[A7 == max(A7)])
mode_qsec
A8<-table(mtcars$vs)
mode_vs <- c(names(A8)[A8 == max(A8)])
mode_vs
A9<-table(mtcars$am)
mode_am <- c(names(A9)[A9 == max(A9)])
mode_am
A10<-table(mtcars$gear)
mode_gear <- c(names(A10)[A10 == max(A10)])
mode_gear
A11<-table(mtcars$carb)
mode_carb <- c(names(A11)[A11 == max(A11)])
mode_carb
```

'A1 <- table(mtcars\$mpg)',
Creates a frequency table of the
'mtcars' dataset's 'mpg' column
values, by counting how many
times each unique value occurs.

'mode_mpg <- c(names(A1)[A1
== max(A1)])', this code
Identifies the 'mpg' column
values that appear most
frequently by finding the names
of mpg values corresponding to
the maximum frequency in the
table.

'mode_mpg', This code displays
the highest frequency values in
the 'mpg' column.

```
> #Mode
> A1 <- table(mtcars$mpg)
> mode_mpg <- c(names(A1)[A1 == max(A1)])
> mode_mpg
[1] "10.4" "15.2" "19.2" "21" "21.4" "22.8" "30.4"
> A2<-table(mtcars$cyl)
> mode_cyl <- c(names(A2)[A2 == max(A2)])
> mode_cyl
[1] "8"
> A3<-table(mtcars$disp)
> mode_disp <- c(names(A3)[A3 == max(A3)])
> mode_disp
[1] "275.8"
> A4<-table(mtcars$hp)
> mode_hp <- c(names(A4)[A4 == max(A4)])
> mode_hp
[1] "110" "175" "180"
> A5<-table(mtcars$drat)
> mode_drat <- c(names(A5)[A5 == max(A5)])
> mode_drat
[1] "3.07" "3.92"
> A6<-table(mtcars$wt)
> mode_wt <- c(names(A6)[A6 == max(A6)])
> mode_wt
[1] "3.44"
> A7<-table(mtcars$qsec)
> mode_qsec <- c(names(A7)[A7 == max(A7)])
> mode_qsec
[1] "17.02" "18.9"
> A8<-table(mtcars$vs)
> mode_vs <- c(names(A8)[A8 == max(A8)])
> mode_vs
[1] "0"
> A9<-table(mtcars$am)
> mode_am <- c(names(A9)[A9 == max(A9)])
> mode_am
[1] "0"
```

```
> A10<-table(mtcars$gear)
> mode_gear <- c(names(A10)[A10 == max(A10)])
> mode_gear
[1] "3"
> A11<-table(mtcars$carb)
> mode_carb <- c(names(A11)[A11 == max(A11)])
> mode_carb
[1] "2" "4"
> |
```

Dispersion statistics

	Code	Output
1. Variance	<pre>#Measures of Dispersion #Variance var(mtcars\$mpg) var(mtcars\$cyl) var(mtcars\$disp) var(mtcars\$hp) var(mtcars\$drat) var(mtcars\$wt) var(mtcars\$qsec) var(mtcars\$vs) var(mtcars\$am) var(mtcars\$gear) var(mtcars\$carb)</pre> <p>Here, 'var(mtcars\$mpg)' calculates the variance of the 'mpg' column in the 'mtcars' dataset 'var()' function in RStudio.</p>	<pre>> #Variance > var(mtcars\$mpg) [1] 36.3241 > var(mtcars\$cyl) [1] 3.189516 > var(mtcars\$disp) [1] 15360.8 > var(mtcars\$hp) [1] 4700.867 > var(mtcars\$drat) [1] 0.2858814 > var(mtcars\$wt) [1] 0.957379 > var(mtcars\$qsec) [1] 3.193166 > var(mtcars\$vs) [1] 0.2540323 > var(mtcars\$am) [1] 0.2489919 > var(mtcars\$gear) [1] 0.5443548 > var(mtcars\$carb) [1] 2.608871 ></pre>
2. Standard deviation	<pre>#Standard_Deviation(sd) sd(mtcars\$mpg) sd(mtcars\$cyl) sd(mtcars\$disp) sd(mtcars\$hp) sd(mtcars\$drat) sd(mtcars\$wt) sd(mtcars\$qsec) sd(mtcars\$vs) sd(mtcars\$am) sd(mtcars\$gear) sd(mtcars\$carb)</pre> <p>Here, 'sd(mtcars\$mpg)' calculates the standard deviation of the 'mpg' column in the 'mtcars' dataset by using 'sd()' function in RStudio.</p>	<pre>> #Standard_Deviation(sd) > sd(mtcars\$mpg) [1] 6.026948 > sd(mtcars\$cyl) [1] 1.785922 > sd(mtcars\$disp) [1] 123.9387 > sd(mtcars\$hp) [1] 68.56287 > sd(mtcars\$drat) [1] 0.5346787 > sd(mtcars\$wt) [1] 0.9784574 > sd(mtcars\$qsec) [1] 1.786943 > sd(mtcars\$vs) [1] 0.5040161 > sd(mtcars\$am) [1] 0.4989909 > sd(mtcars\$gear) [1] 0.7378041 > sd(mtcars\$carb) [1] 1.6152 ></pre>

3. Minimum Value

```
#Minimum_Value(min)
min_mpg<-min(mtcars$mpg)
min_mpg
min_cyl<-min(mtcars$cyl)
min_cyl
min_disp<-min(mtcars$disp)
min_disp
min_hp<-min(mtcars$hp)
min_hp
min_drat<-min(mtcars$drat)
min_drat
min_wt<-min(mtcars$wt)
min_wt
min_qsec<-min(mtcars$qsec)
min_qsec
min_vs<-min(mtcars$vs)
min_vs
min_am<-min(mtcars$am)
min_am
min_gear<-min(mtcars$gear)
min_gear
min_carb<-min(mtcars$carb)
min_carb
```

The 'min_mpg <- min(mtcars\$mpg)', line calculates the minimum mpg value in the mtcars dataset by using the 'min()' function and stores it in the variable 'min_mpg'.

The 'min_mpg', line outputs the value stored in min_mpg.

```
> #Minimum_Value(min)
> min_mpg<-min(mtcars$mpg)
> min_mpg
[1] 10.4
> min_cyl<-min(mtcars$cyl)
> min_cyl
[1] 4
> min_disp<-min(mtcars$disp)
> min_disp
[1] 71.1
> min_hp<-min(mtcars$hp)
> min_hp
[1] 52
> min_drat<-min(mtcars$drat)
> min_drat
[1] 2.76
> min_wt<-min(mtcars$wt)
> min_wt
[1] 1.513
> min_qsec<-min(mtcars$qsec)
> min_qsec
[1] 14.5
> min_vs<-min(mtcars$vs)
> min_vs
[1] 0
> min_am<-min(mtcars$am)
> min_am
[1] 0
> min_gear<-min(mtcars$gear)
> min_gear
[1] 3
> min_carb<-min(mtcars$carb)
> min_carb
[1] 1
>
```

4. Maximum Value

```
#Maximum_Value(max)
max_mpg<-max(mtcars$mpg)
max_mpg
max_cyl<-max(mtcars$cyl)
max_cyl
max_disp<-max(mtcars$disp)
max_disp
max_hp<-max(mtcars$hp)
max_hp
max_drat<-max(mtcars$drat)
max_drat
max_wt<-max(mtcars$wt)
max_wt
max_qsec<-max(mtcars$qsec)
max_qsec
max_vs<-max(mtcars$vs)
max_vs
max_am<-max(mtcars$am)
max_am
max_gear<-max(mtcars$gear)
max_gear
max_carb<-max(mtcars$carb)
max_carb
```

The 'max_mpg <- max(mtcars\$mpg)', line calculates the maximum mpg value in the mtcars dataset by using the 'max()' function and stores it in the variable 'max_mpg'.

The 'max_mpg', line outputs the value stored in 'max_mpg'.

```
> #Maximum_Value(max)
> max_mpg<-max(mtcars$mpg)
> max_mpg
[1] 33.9
> max_cyl<-max(mtcars$cyl)
> max_cyl
[1] 8
> max_disp<-max(mtcars$disp)
> max_disp
[1] 472
> max_hp<-max(mtcars$hp)
> max_hp
[1] 335
> max_drat<-max(mtcars$drat)
> max_drat
[1] 4.93
> max_wt<-max(mtcars$wt)
> max_wt
[1] 5.424
> max_qsec<-max(mtcars$qsec)
> max_qsec
[1] 22.9
> max_vs<-max(mtcars$vs)
> max_vs
[1] 1
> max_am<-max(mtcars$am)
> max_am
[1] 1
> max_gear<-max(mtcars$gear)
> max_gear
[1] 5
> max_carb<-max(mtcars$carb)
> max_carb
[1] 8
>
```

<p>5. Range</p>	<pre>#Range r_mpg<-c(max_mpg-min_mpg) r_mpg r_cyl<-c(max_cyl-min_cyl) r_cyl r_disp<-c(max_disp-min_disp) r_disp r_hp<-c(max_hp-min_hp) r_hp r_drat<-c(max_drat-min_drat) r_drat r_wt<-c(max_wt-min_wt) r_wt r_qsec<-c(max_qsec-min_qsec) r_qsec r_vs<-c(max_vs-min_vs) r_vs r_am<-c(max_am-min_am) r_am r_gear<-c(max_gear-min_gear) r_gear r_carb<-c(max_carb-min_carb) r_carb</pre> <p>The 'r_mpg<-c(max_mpg-min_mpg)', line calculate the range by deducting maximum mpg (max_mpg) from minimum mpg (min_mpg) and it in the variable 'r_mpg'.</p> <p>The 'r_mpg', line outputs the value stored in 'r_mpg'.</p>	<pre>> #Range > r_mpg<-c(max_mpg-min_mpg) > r_mpg [1] 23.5 > r_cyl<-c(max_cyl-min_cyl) > r_cyl [1] 4 > r_disp<-c(max_disp-min_disp) > r_disp [1] 400.9 > r_hp<-c(max_hp-min_hp) > r_hp [1] 283 > r_drat<-c(max_drat-min_drat) > r_drat [1] 2.17 > r_wt<-c(max_wt-min_wt) > r_wt [1] 3.911 > r_qsec<-c(max_qsec-min_qsec) > r_qsec [1] 8.4 > r_vs<-c(max_vs-min_vs) > r_vs [1] 1 > r_am<-c(max_am-min_am) > r_am [1] 1 > r_gear<-c(max_gear-min_gear) > r_gear [1] 2 > r_carb<-c(max_carb-min_carb) > r_carb [1] 7 ></pre>
<p>6. Interquartile Range</p>	<pre>#Interquartile_Range(IQR) IQR(mtcars\$mpg) IQR(mtcars\$cyl) IQR(mtcars\$disp) IQR(mtcars\$hp) IQR(mtcars\$drat) IQR(mtcars\$wt) IQR(mtcars\$qsec) IQR(mtcars\$vs) IQR(mtcars\$am) IQR(mtcars\$gear) IQR(mtcars\$carb)</pre> <p>Here, 'IQR(mtcars\$mpg)' calculates the standard deviation of the 'mpg' column in the 'mtcars' dataset by using 'sd()' function in RStudio.</p>	<pre>> #Interquartile_Range(IQR) > IQR(mtcars\$mpg) [1] 7.375 > IQR(mtcars\$cyl) [1] 4 > IQR(mtcars\$disp) [1] 205.175 > IQR(mtcars\$hp) [1] 83.5 > IQR(mtcars\$drat) [1] 0.84 > IQR(mtcars\$wt) [1] 1.02875 > IQR(mtcars\$qsec) [1] 2.0075 > IQR(mtcars\$vs) [1] 1 > IQR(mtcars\$am) [1] 1 > IQR(mtcars\$gear) [1] 1 > IQR(mtcars\$carb) [1] 2 ></pre>

3.3

- i) How many plots are possible in theory? How many scatter plots are there, if one uses each variable pair once? Use factorial-based formulas.

$${}^nC_r = n! / (r! (n - r)!)$$

$n = 4$ (number of numerical variables)

$r = 2$ (number of variables pair at a time)

$${}^4C_2 = 4! / (2! (4 - 2)!)$$

$$= 4 * 3 * 2 * 1 / (2! (2)!)$$

$$= 24 / (2 * 1 * 2 * 1)$$

$$= 24 / 4$$

$$= 6$$

In theory it's possible to have 6 plots.

Code

```
#Q1
num_variables<-4
num_pair<-2
Number_of_scatter_plots<-combn(num_variables, num_pair)
Number_of_scatter_plots
```

Here use 'combn()' function in RStudio to calculate the combination for this question.

Output

```
> #Question_3.3
> #Q1
> num_variables<-4
> num_pair<-2
> Number_of_scatter_plots<-combn(num_variables, num_pair)
> Number_of_scatter_plots
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    1    1    2    2    3
[2,]    2    3    4    3    4    4
> |
```

ii) Plot all scatter plots using each variable pair at once.

Code

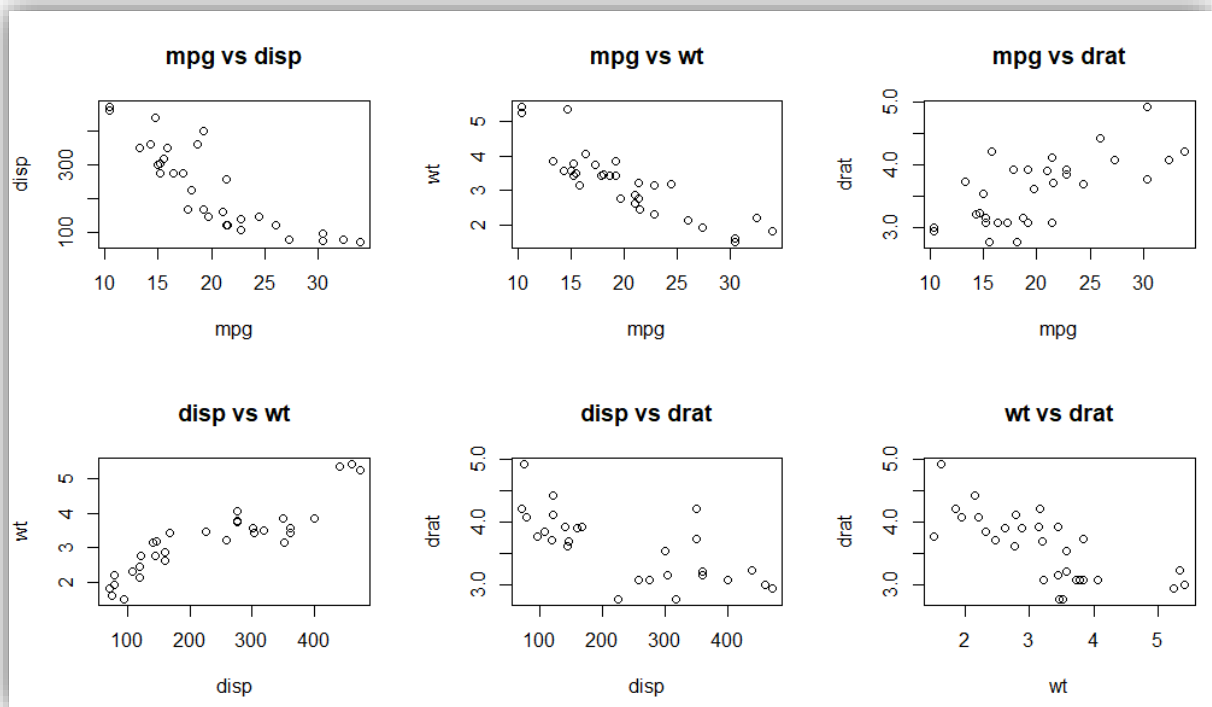
```
#Q2
#Plotting scatter plots
par(mfrow = c(3,3))
plot(mtcars$mpg, mtcars$disp,main = "mpg vs disp",xlab = "mpg",ylab = "disp")
plot(mtcars$mpg,mtcars$wt,main="mpg vs wt",xlab="mpg",ylab="wt")
plot(mtcars$mpg, mtcars$drat,main = "mpg vs drat",xlab = "mpg",ylab = "drat")
plot(mtcars$disp, mtcars$wt,main = "disp vs wt",xlab = "disp",ylab = "wt")
plot(mtcars$disp, mtcars$drat,main = "disp vs drat",xlab = "disp",ylab = "drat")
plot(mtcars$wt, mtcars$drat,main = "wt vs drat",xlab = "wt",ylab = "drat")
```

The 'par(mfrow = c(3, 3))' line divide the plotting area into a 3x3 grid, allowing six plots to be displayed in one frame.

The 'plot(mtcars\$mpg, mtcars\$disp, main = "mpg vs disp", xlab = "mpg", ylab = "disp")' code generates a scatter plot to visualize the relationship between 'mpg' and 'disp' from the 'mtcars' dataset. Here 'plot()' function is used to create a scatter plot by plotting the 'mpg' values on the x axis and the 'disp' values on the y axis. The 'main' is used for giving a title to plot, which is set as 'mpg vs disp'. The 'xlab' sets the label for the x axis as 'mpg' and the 'ylab' sets the label for the y axis as 'disp',

Output

```
> #Q2
> #Plotting scatter plots
> par(mfrow = c(3,3))
> plot(mtcars$mpg, mtcars$disp,main = "mpg vs disp",xlab = "mpg",ylab = "disp")
> plot(mtcars$mpg,mtcars$wt,main="mpg vs wt",xlab="mpg",ylab="wt")
> plot(mtcars$mpg, mtcars$drat,main = "mpg vs drat",xlab = "mpg",ylab = "drat")
> plot(mtcars$disp, mtcars$wt,main = "disp vs wt",xlab = "disp",ylab = "wt")
> plot(mtcars$disp, mtcars$drat,main = "disp vs drat",xlab = "disp",ylab = "drat")
> plot(mtcars$wt, mtcars$drat,main = "wt vs drat",xlab = "wt",ylab = "drat")
```

iii) Indicate which plots seem to exhibit a linear relationship? Express in one sentence what such a relationship indicates.

- By looking at all the plots, other than 'disp vs drat', all other five plots ('mpg vs disp', 'mpg vs wt', 'mpg vs drat', 'disp vs wt' and 'wt vs drat') shows a linear relationship. That because only those 5 plots resemble a straight line either upward or downward.
- This type of relationship indicates that when one variable changes, the other variable tends to change as well.

3.4

- i) Create (relative) frequency histograms for each continuous variable.

continuous variables of the mtcars dataset are,

'mpg', 'disp', 'hp', 'drat', 'wt', 'qsec'

Code

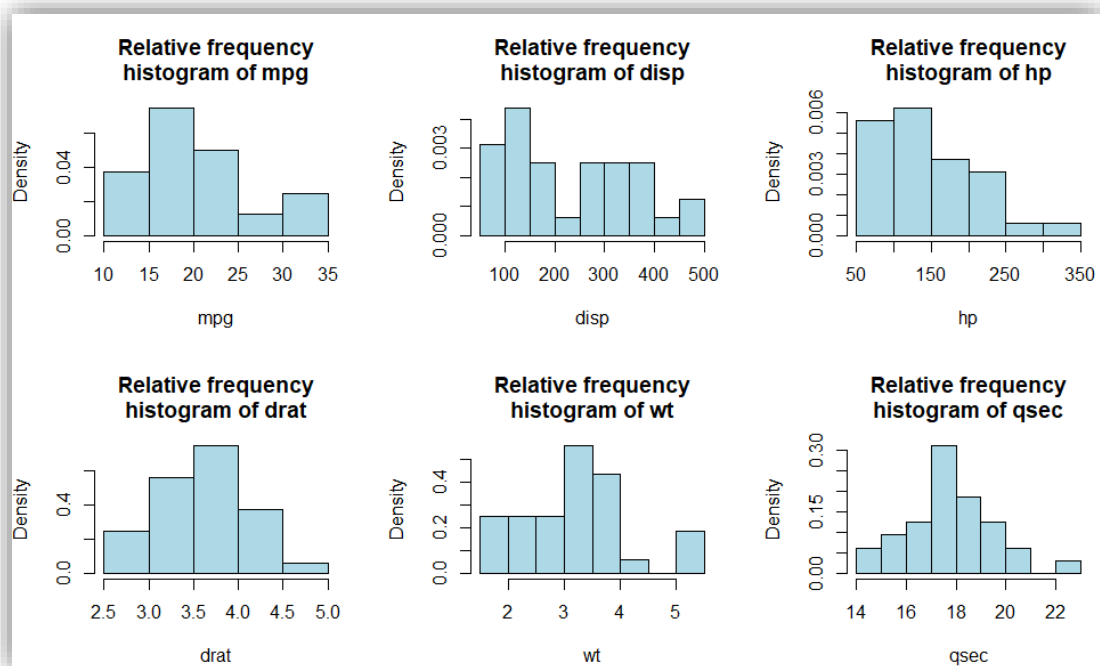
```
#Question_3.4
#Q1
#Histogram for each continuous variable
par(mfrow = c(3,3))
hist(mtcars$mpg,freq=FALSE,main="Relative frequency
histogram of mpg",xlab="mpg",col="lightblue")
hist(mtcars$disp,freq=FALSE,main="Relative frequency
histogram of disp",xlab="disp",col="lightblue")
hist(mtcars$hp,freq=FALSE,main="Relative frequency
histogram of hp",xlab="hp",col="lightblue")
hist(mtcars$drat,freq=FALSE,main="Relative frequency
histogram of drat",xlab="drat",col="lightblue")
hist(mtcars$wt,freq=FALSE,main="Relative frequency
histogram of wt",xlab="wt",col="lightblue")
hist(mtcars$qsec,freq=FALSE,main="Relative frequency
histogram of qsec",xlab="qsec",col="lightblue")
```

The 'par(mfrow = c(3, 3))' line divide the plotting area into a 3x3 grid, allowing six plots to be displayed in one frame.

'hist(mtcars\$mpg,freq=FALSE,main="Relative frequency histogram of mpg",col="lightblue")', The line generates a histogram for the 'mpg' column from the 'mtcars' dataset by using 'hist()' function in R. Here, by setting the 'freq' argument to 'FALSE', the histogram displays the density. The 'main' argument makes the title of the plot as 'Relative frequency histogram of mpg'. The 'xlab = "mpg"' labels the x axis as 'mpg'. Also, the 'col' argument sets the color of the bars to 'lightblue'.

Output

```
> #Question_3.4
> #Q1
> #Histogram for each continuous variable
> par(mfrow = c(3,3))
> hist(mtcars$mpg,freq=FALSE,main="Relative frequency
+ histogram of mpg",xlab="mpg",col="lightblue")
> hist(mtcars$disp,freq=FALSE,main="Relative frequency
+ histogram of disp",xlab="disp",col="lightblue")
> hist(mtcars$hp,freq=FALSE,main="Relative frequency
+ histogram of hp",xlab="hp",col="lightblue")
> hist(mtcars$drat,freq=FALSE,main="Relative frequency
+ histogram of drat",xlab="drat",col="lightblue")
> hist(mtcars$wt,freq=FALSE,main="Relative frequency
+ histogram of wt",xlab="wt",col="lightblue")
> hist(mtcars$qsec,freq=FALSE,main="Relative frequency
+ histogram of qsec",xlab="qsec",col="lightblue")
```



ii) Which variables appear to be normally distributed? Justify your answer.

- Among the variables in the dataset 'mpg', 'drat', 'wt', and 'qsec' are all normally distributed.
- Justification

Code

```
#Q2
#curve
par(mfrow = c(3,4))
hist(mtcars$mpg,freq=FALSE,main="Curve of mpg",xlab="mpg")
curve(dnorm(x,mean(mtcars$mpg),sd(mtcars$mpg)),add=TRUE,col="red")
qqnorm(mtcars$mpg,main="Q-Q Plot of mpg")
qqline(mtcars$mpg,col="red")

hist(mtcars$drat,freq=FALSE,main="Curve of drat",xlab="drat")
curve(dnorm(x,mean(mtcars$drat),sd(mtcars$drat)),add=TRUE,col="red")
qqnorm(mtcars$drat,main="Q-Q Plot of drat")
qqline(mtcars$drat,col="red")

hist(mtcars$wt,freq=FALSE,main="Curve of wt",xlab="wt")
curve(dnorm(x,mean(mtcars$wt),sd(mtcars$wt)),add=TRUE,col="red")
qqnorm(mtcars$wt,main="Q-Q Plot of wt")
qqline(mtcars$wt,col="red")

hist(mtcars$qsec,freq=FALSE,main="Curve of qsec",xlab="qsec")
curve(dnorm(x,mean(mtcars$qsec),sd(mtcars$qsec)),add=TRUE,col="red")
qqnorm(mtcars$qsec,main="Q-Q Plot of qsec")
qqline(mtcars$qsec,col="red")

hist(mtcars$disp,freq=FALSE,main="Curve of disp",xlab="disp")
curve(dnorm(x,mean(mtcars$disp),sd(mtcars$disp)),add=TRUE,col="red")
qqnorm(mtcars$disp,main="Q-Q Plot of disp")
qqline(mtcars$disp,col="red")

hist(mtcars$hp,freq=FALSE,main="Curve of hp",xlab="hp")
curve(dnorm(x,mean(mtcars$hp),sd(mtcars$hp)),add=TRUE,col="red")
qqnorm(mtcars$hp,main="Q-Q Plot of hp")
qqline(mtcars$hp,col="red")
```

The 'par(mfrow = c(3, 4))' line divide the plotting area into a 3x3 grid, allowing six plots to be displayed in one frame.

The line 'hist(mtcars\$mpg,freq=FALSE,main="Curve of mpg",xlab="mpg")' generates a histogram for the 'mpg' column from the 'mtcars' dataset by using 'hist()' function in R. Here, by setting the 'freq' argument to 'FALSE', the histogram displays the density. The 'main' argument makes the title of the plot as 'Curve of mpg'. The 'xlab = "mpg"' labels the x axis as 'mpg'.

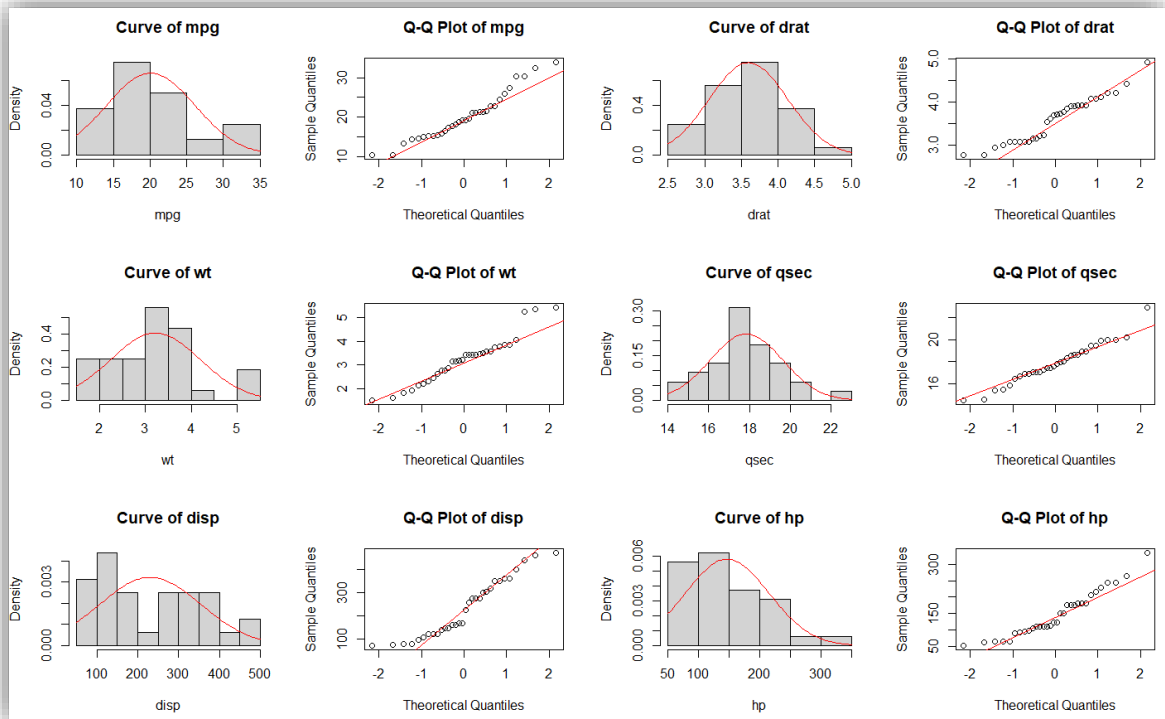
The line `curve(dnorm(x, mean(mtcars$mpg), sd(mtcars$mpg)), add=TRUE, col="red")`, add a normal distribution curve to the existing histogram. The function `dnorm(x, mean = mean(mtcars$mpg), sd = sd(mtcars$mpg))` calculates the values of normal distribution using the mean and standard deviation of the 'mpg' variable from the 'mtcars' dataset. By setting `'add = TRUE'`, the curve is added onto the previously generated histogram. Also, the `'col'` argument sets the color of the curve to 'red'.

The line `qqnorm(mtcars$mpg,main="Q-Q Plot of mpg")`, creates a Q-Q (quantile-quantile) plot for the 'mpg' variable from the 'mtcars' dataset. The `'main'` argument makes the title of the plot as 'Q-Q Plot of mpg'.

The line, `qqline(mtcars$mpg,col="red")`, adds a diagonal reference line to the Q-Q plot, representing the expected relationship if the data were perfectly normally distributed. This is done by using `'qqline'` function in RStudio. the `'col'` argument sets the color of the reference line to 'red'

Output

```
> #curve
> par(mfrow = c(3,4))
> hist(mtcars$mpg,freq=FALSE,main="Curve of mpg",xlab="mpg")
> curve(dnorm(x,mean(mtcars$mpg),sd(mtcars$mpg)),add=TRUE,col="red")
> qqnorm(mtcars$mpg,main="Q-Q Plot of mpg")
> qqline(mtcars$mpg,col="red")
> hist(mtcars$drat,freq=FALSE,main="Curve of drat",xlab="drat")
> curve(dnorm(x,mean(mtcars$drat),sd(mtcars$drat)),add=TRUE,col="red")
> qqnorm(mtcars$drat,main="Q-Q Plot of drat")
> qqline(mtcars$drat,col="red")
> hist(mtcars$wt,freq=FALSE,main="Curve of wt",xlab="wt")
> curve(dnorm(x,mean(mtcars$wt),sd(mtcars$wt)),add=TRUE,col="red")
> qqnorm(mtcars$wt,main="Q-Q Plot of wt")
> qqline(mtcars$wt,col="red")
> hist(mtcars$qsec,freq=FALSE,main="Curve of qsec",xlab="qsec")
> curve(dnorm(x,mean(mtcars$qsec),sd(mtcars$qsec)),add=TRUE,col="red")
> qqnorm(mtcars$qsec,main="Q-Q Plot of qsec")
> qqline(mtcars$qsec,col="red")
> hist(mtcars$disp,freq=FALSE,main="Curve of disp",xlab="disp")
> curve(dnorm(x,mean(mtcars$disp),sd(mtcars$disp)),add=TRUE,col="red")
> qqnorm(mtcars$disp,main="Q-Q Plot of disp")
> qqline(mtcars$disp,col="red")
> hist(mtcars$hp,freq=FALSE,main="Curve of hp",xlab="hp")
> curve(dnorm(x,mean(mtcars$hp),sd(mtcars$hp)),add=TRUE,col="red")
> qqnorm(mtcars$hp,main="Q-Q Plot of hp")
> qqline(mtcars$hp,col="red")
```



The bell-shaped histograms with normal distribution curves suggest that the variables are normally distributed. This visual indication is further supported by Q-Q plots, where the data points align closely with the diagonal reference line, indicating that the variances correspond to a normal distribution. These combined observations strengthen the assumption of normality of the abovementioned variables.

- iii) For each variable that is confirmed to have a normal distribution, based on its mean and its standard deviation, define (in a formula) the model normal probability density function and plot it. Compare the theoretical plot with the histogram.

Code

```
#Q3
par(mfrow = c(2,2))

# mpg
mpg<-mtcars$mpg
mpg
mpg_mean<-mean(mpg)
mpg_mean
mpg_sd<-sd(mpg)
mpg_sd
# Curve
hist(mpg, freq=FALSE, main="Curve of mpg", xlab="mpg")
curve(dnorm(x, mean=mpg_mean, sd=mpg_sd), add=TRUE, col="red")

# drat
drat<-mtcars$drat
drat
drat_mean<-mean(drat)
drat_mean
drat_sd<-sd(drat)
drat_sd
# Curve
hist(drat, freq=FALSE, main="Curve of drat", xlab="drat")
curve(dnorm(x, mean=drat_mean, sd=drat_sd), add=TRUE, col="red")

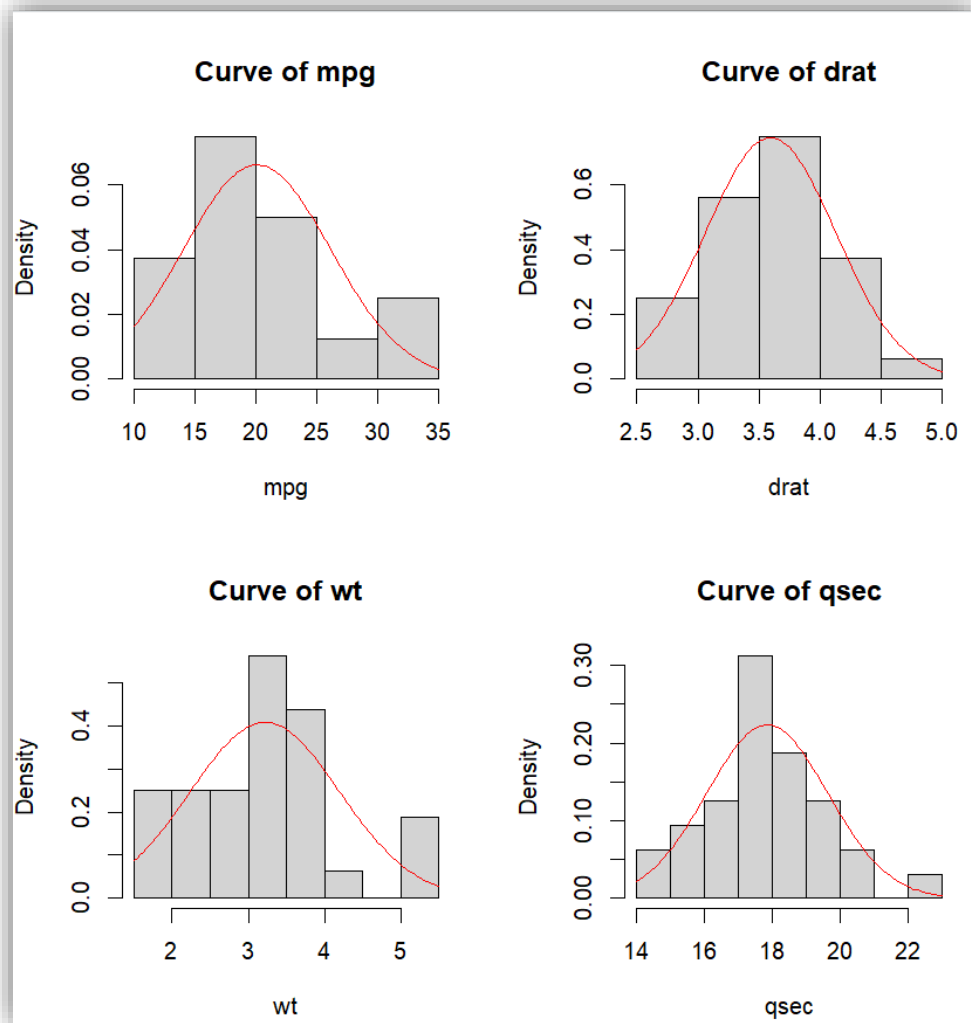
# wt
wt<-mtcars$wt
wt
wt_mean<-mean(wt)
wt_mean
wt_sd<-sd(wt)
wt_sd
# Curve
hist(wt, freq=FALSE, main="Curve of wt", xlab="wt")
curve(dnorm(x, mean=wt_mean, sd=wt_sd), add=TRUE, col="red")
```

```
# qsec
qsec<-mtcars$qsec
qsec
qsec_mean<-mean(qsec)
qsec_mean
qsec_sd<-sd(qsec)
qsec_sd
# Curve
hist(qsec, freq=FALSE, main="Curve of qsec", xlab="qsec")
curve(dnorm(x, mean=qsec_mean, sd=qsec_sd), add=TRUE, col="red")
```

Output

```
> #Q3
> par(mfrow = c(2,2))
> # mpg
> mpg<-mtcars$mpg
> mpg
[1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5
[23] 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4
> mpg_mean<-mean(mpg)
> mpg_mean
[1] 20.09062
> mpg_sd<-sd(mpg)
> mpg_sd
[1] 6.026948
> # Curve
> hist(mpg, freq=FALSE, main="Curve of mpg", xlab="mpg")
> curve(dnorm(x, mean=mpg_mean, sd=mpg_sd), add=TRUE, col="red")
> # drat
> drat<-mtcars$drat
> drat
[1] 3.90 3.90 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 3.92 3.07 3.07 3.07 2.93 3.00 3.23 4.08 4.93 4.22 3.70 2.76
[23] 3.15 3.73 3.08 4.08 4.43 3.77 4.22 3.62 3.54 4.11
> drat_mean<-mean(drat)
> drat_mean
[1] 3.596563
> drat_sd<-sd(drat)
> drat_sd
[1] 0.5346787
> # Curve
> hist(drat, freq=FALSE, main="Curve of drat", xlab="drat")
> curve(dnorm(x, mean=drat_mean, sd=drat_sd), add=TRUE, col="red")
```

```
> # wt
> wt<-mtcars$wt
> wt
[1] 2.620 2.875 2.320 3.215 3.440 3.460 3.570 3.190 3.150 3.440 3.440 4.070 3.730 3.780 5.250 5.424 5.345 2.200
[19] 1.615 1.835 2.465 3.520 3.435 3.840 3.845 1.935 2.140 1.513 3.170 2.770 3.570 2.780
> wt_mean<-mean(wt)
> wt_mean
[1] 3.21725
> wt_sd<-sd(wt)
> wt_sd
[1] 0.9784574
> # Curve
> hist(wt, freq=FALSE, main="Curve of wt", xlab="wt")
> curve(dnorm(x, mean=wt_mean, sd=wt_sd), add=TRUE, col="red")
> # qsec
> qsec<-mtcars$qsec
> qsec
[1] 16.46 17.02 18.61 19.44 17.02 20.22 15.84 20.00 22.90 18.30 18.90 17.40 17.60 18.00 17.98 17.82 17.42 19.47
[19] 18.52 19.90 20.01 16.87 17.30 15.41 17.05 18.90 16.70 16.90 14.50 15.50 14.60 18.60
> qsec_mean<-mean(qsec)
> qsec_mean
[1] 17.84875
> qsec_sd<-sd(qsec)
> qsec_sd
[1] 1.786943
> # Curve
> hist(qsec, freq=FALSE, main="Curve of qsec", xlab="qsec")
> curve(dnorm(x, mean=qsec_mean, sd=qsec_sd), add=TRUE, col="red")
```

The normal probability density function (PDF) was defined for each of these variables based on their respective means and standard deviations. A curve representing the theoretical normal distribution was overlaid on each histogram. The comparison between the theoretical normal distribution and the observed data distribution showed that the histograms of 'mpg', 'drat', 'wt' and 'qsec' closely followed the theoretical normal curves, confirming their approximate normality.

3.5

- i) $X < 250$
- ii) $230 < X < 250$
- iii) $X < 220$

Code

```
#Question_3.5
#disp
#X = disp
X_mean<-mean(mtcars$disp)
X_sd<-sd(mtcars$disp)

#Q1
#X>250
pnorm(250,X_mean,X_sd,lower.tail = FALSE)
#percentage
pnorm(250,X_mean,X_sd,lower.tail = FALSE)*100

#Q2
#230<X<250
pnorm(250,X_mean,X_sd,lower.tail = TRUE)-pnorm(230,X_mean,X_sd,lower.tail = TRUE)
#percentage
(pnorm(250,X_mean,X_sd,lower.tail = TRUE)-pnorm(230,X_mean,X_sd,lower.tail = TRUE))*100

#Q3
#X<220
pnorm(220,X_mean,X_sd,lower.tail = TRUE)
#percentage
pnorm(220,X_mean,X_sd,lower.tail = TRUE)*100
```

Here by using the 'pnorm' function, this R code computes probability for the 'disp' variable from the 'mtcars' dataset.

Output

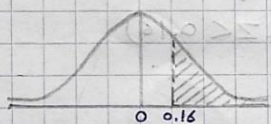
```
> #Question_3.5
> #disp
> #X = disp
> X_mean<-mean(mtcars$disp)
> X_sd<-sd(mtcars$disp)
> #Q1
> #X>250
> pnorm(250,X_mean,X_sd,lower.tail = FALSE)
[1] 0.4381956
> #percentage
> pnorm(250,X_mean,X_sd,lower.tail = FALSE)*100
[1] 43.81956
> #Q2
> #230<X<250
> pnorm(250,X_mean,X_sd,lower.tail = TRUE)-pnorm(230,X_mean,X_sd,lower.tail = TRUE)
[1] 0.06412802
> #percentage
> (pnorm(250,X_mean,X_sd,lower.tail = TRUE)-pnorm(230,X_mean,X_sd,lower.tail = TRUE))*100
[1] 6.412802
> #Q3
> #X<220
> pnorm(220,X_mean,X_sd,lower.tail = TRUE)
[1] 0.4655307
> #percentage
> pnorm(220,X_mean,X_sd,lower.tail = TRUE)*100
[1] 46.55307
```

By hand;

CW - Refer Defer

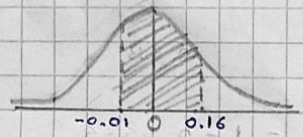
5) Mean / $\mu = 230.72$ $sd / \sigma = 123.94$

i) $x > 250$

$$z = \frac{x - \mu}{\sigma}$$
$$P\left(z > \frac{250 - \mu}{\sigma}\right) = P\left(z > \frac{250 - 230.72}{123.94}\right)$$
$$= P(z > 0.16)$$

$$P(z > 0.16) = 1 - 0.5636$$
$$= 0.4364$$
$$P(z > 0.16) = 0.4364 \times 100\%$$
$$= 43.64\%$$

5) Mean / $\mu = 230.72$ $sd / \sigma = 123.94$

ii) $230 < x < 250$

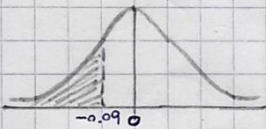
$$z = \frac{x - \mu}{\sigma}$$
$$P\left(\frac{230 - \mu}{\sigma} < z < \frac{250 - \mu}{\sigma}\right)$$
$$= P\left(\frac{230 - 230.72}{123.94} < z < \frac{250 - 230.72}{123.94}\right)$$
$$= P(-0.01 < z < 0.16)$$

$$P(-0.01 < z < 0.16) = 0.5636 - (1 - 0.504)$$
$$= 0.0676$$
$$P(-0.01 < z < 0.16) = 0.0676 \times 100\%$$
$$= 6.76\%$$

5) Mean / $\mu = 230.72$ $sd / \sigma = 123.94$

ii) $X < 220$

$$Z = \frac{X - \mu}{\sigma}$$

$$P\left(Z < \frac{220 - \mu}{\sigma}\right) = P\left(Z < \frac{220 - 230.72}{123.94}\right)$$

$$= P(Z < -0.09)$$


$$P(Z < -0.09) = 1 - 0.5359$$

$$= 0.4641$$

$$P(Z < -0.09) = 0.4641 \times 100\%$$

$$= 46.41\%$$

In this context, the probabilities tell us how likely it is for a car in the 'mtcars' dataset to have a 'displacement' ('disp') that falls within or beyond specific ranges, when assuming the data is normally distributed.

3.6

i)

Code

```
#Question_3.6
#Q1
#X = disp
X_mean<-mean(mtcars$disp)
X_sd<-sd(mtcars$disp)
X_sample_size<-length(mtcars$disp)
CI_95<-qnorm(0.975)

#Determine Confidence interval
lowerCI_X<-X_mean-CI_95*(X_sd/sqrt(X_sample_size))
lowerCI_X
UpperCI_X<-X_mean+CI_95*(X_sd/sqrt(X_sample_size))
UpperCI_X
```

- The 'X_mean<-mean(mtcars\$disp)', line calculates the mean 'disp' value in the 'mtcars' dataset by using the 'mean()' function and stores it in the variable 'X_mean'.
- The 'X_sd<-sd(mtcars\$disp)', line calculates the standard deviation 'disp' value in the 'mtcars' dataset by using the 'sd()' function and stores it in the variable 'X_sd'.
- The 'X_sample_size<-length(mtcars\$disp)', line calculates the length of the 'disp' in the 'mtcars' dataset by using the 'length()' function and stores it in the variable 'X_sample_size'.
- The 'CI_95 <-qnorm(0.975)', line calculates the sample size of the 'disp' value in the 'mtcars' dataset by using the 'qnorm()' function and stores it in the variable 'CI_95'.
- The lines 'lowerCI_X<-X_mean-CI_95*(X_sd/sqrt(X_sample_size))' and 'UpperCI_X<-X_mean+CI_95*(X_sd/sqrt(X_sample_size))', calculates the lower and upper level of the 95% confidence interval for the population mean. In this calculation, the 'CI_95 * (X_sd / sqrt(X_sample_size))' represents the standard error of mean (STEM). This standard error of mean is both added to and subtracted from the sample mean to determine the full confidence interval. By subtracting it from the sample mean, the lower level of the confidence interval is obtained, which gives an estimate of the minimum plausible value of the population mean with 95% confidence and vice versa by adding to the sample mean, the upper level of the confidence interval is obtained, which gives an estimate of the maximum plausible value of the population mean with 95%

confidence. Then this will be stored in variables called 'lowerCI_X' and 'UpperCI_X' respectively.

- The 'lowerCI_X' and 'UpperCI_X', lines output the values stored in them.

Output

```
> #Question_3.6
> #Q1
> #X = disp
> X_mean<-mean(mtcars$disp)
> X_sd<-sd(mtcars$disp)
> X_sample_size<-length(mtcars$disp)
> CI_95<-qnorm(0.975)
> #Determine Confidence interval
> lowerCI_X<-X_mean-CI_95*(X_sd/sqrt(X_sample_size))
> lowerCI_X
[1] 187.7801
> UpperCI_X<-X_mean+CI_95*(X_sd/sqrt(X_sample_size))
> UpperCI_X
[1] 273.6637
```

There is a 95% chance that the true mean displacement of the population of cars lies between 187.78 and 273.67 cubic inches.

ii)

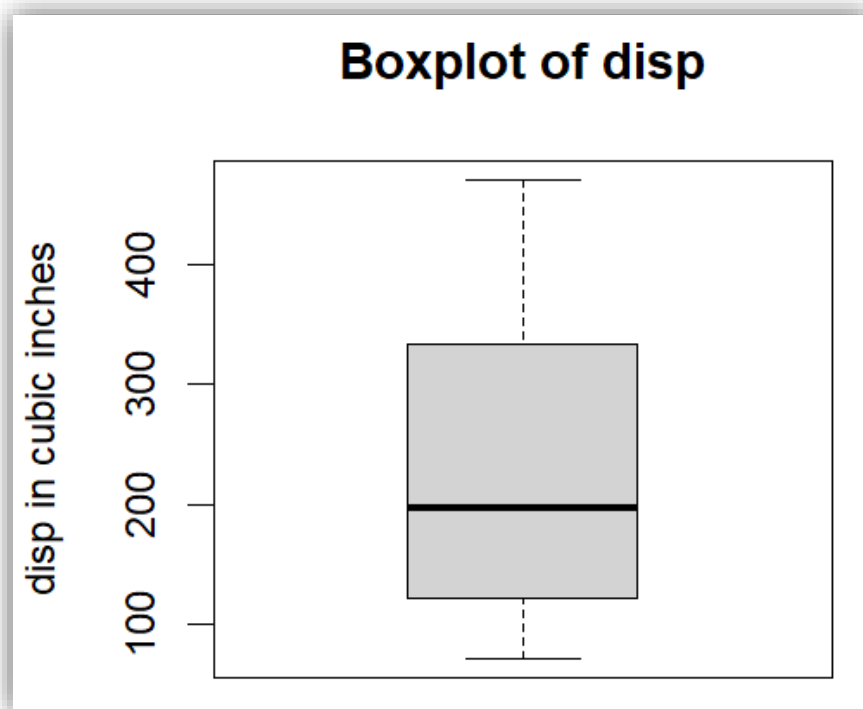
Code

```
#Q2  
boxplot(mtcars$disp,main = "Boxplot of disp",ylab = "disp in cubic inches")
```

The 'boxplot(mtcars\$disp,main = "Boxplot of disp",ylab = "disp in cubic inches")' code generates a box plot to identify the outliers in the 'mtcars' dataset. Here 'boxplot()' function is used to create a box plot by plotting the 'disp' values on the y axis. The 'main' is used for giving a title to plot, which is set as 'Boxplot of disp'. The 'ylab' sets the label for the y axis as 'disp in cubic inches'.

Output

```
#Q2  
boxplot(mtcars$disp,main = "Boxplot of disp",ylab = "disp in cubic inches")
```



By looking at this boxplot we can understand there's no outliers within the 'disp'.