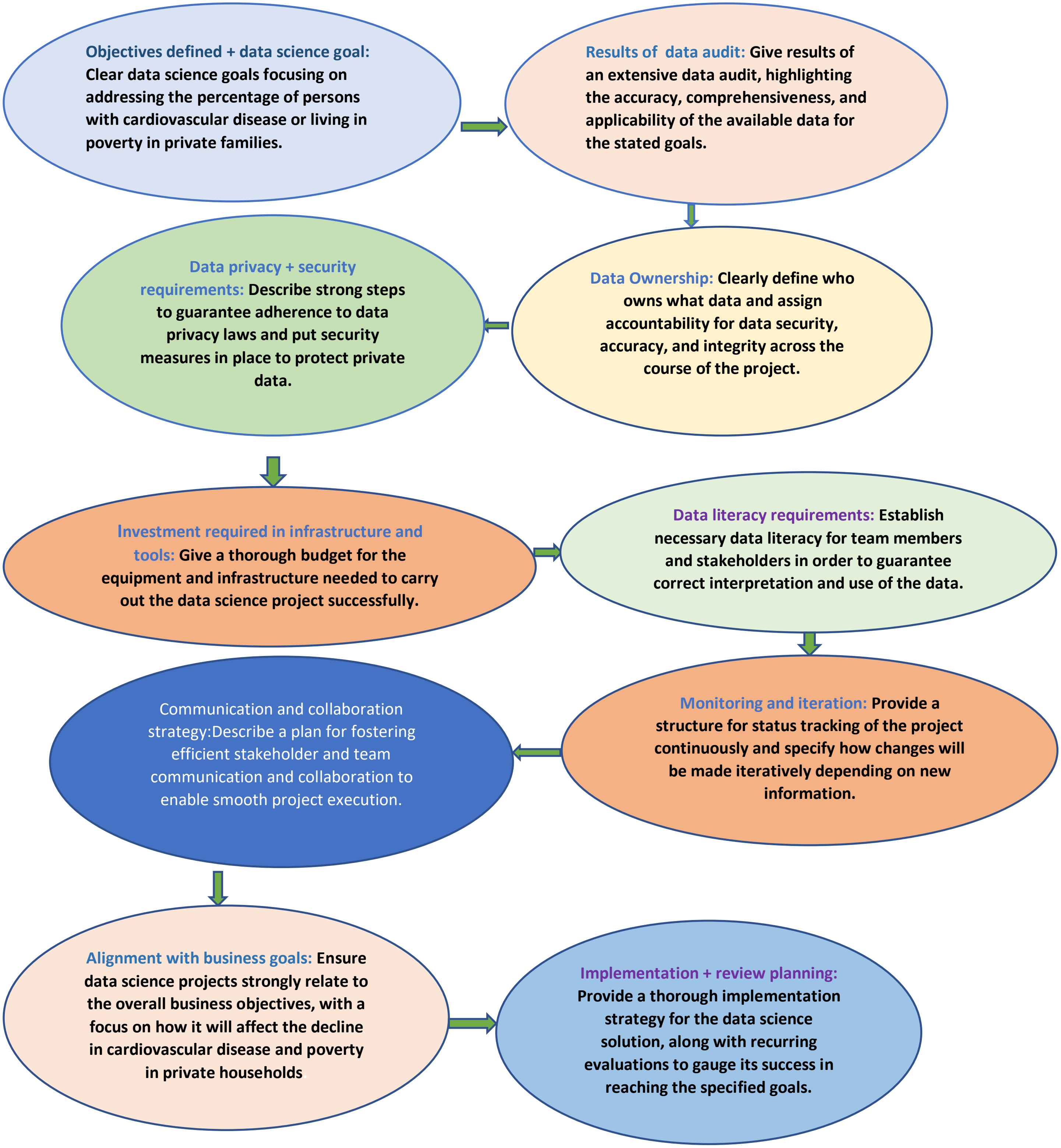
SECTION 1: Estimating the percentage of persons with cardiovascular diseases or living in private households



Objectives defined + data science goal: Clear data science goals focusing on addressing the percentage of persons with cardiovascular disease or living in poverty in private families.

Results of data audit: Give results of an extensive data audit, highlighting the accuracy, comprehensiveness, and applicability of the available data for the stated goals.

Data Ownership: Clearly define who owns what data and assign accountability for data security, accuracy, and integrity across the course of the project.

Data privacy + security requirements: Describe strong steps to guarantee adherence to data privacy laws and put security measures in place to protect private data.

Investment required in infrastructure and tools: Give a thorough budget for the equipment and infrastructure needed to carry out the data science project successfully.

Data literacy requirements: Establish necessary data literacy for team members and stakeholders in order to guarantee correct interpretation and use of the data.

Monitoring and iteration: Provide a structure for status tracking of the project continuously and specify how changes will be made iteratively depending on new information.

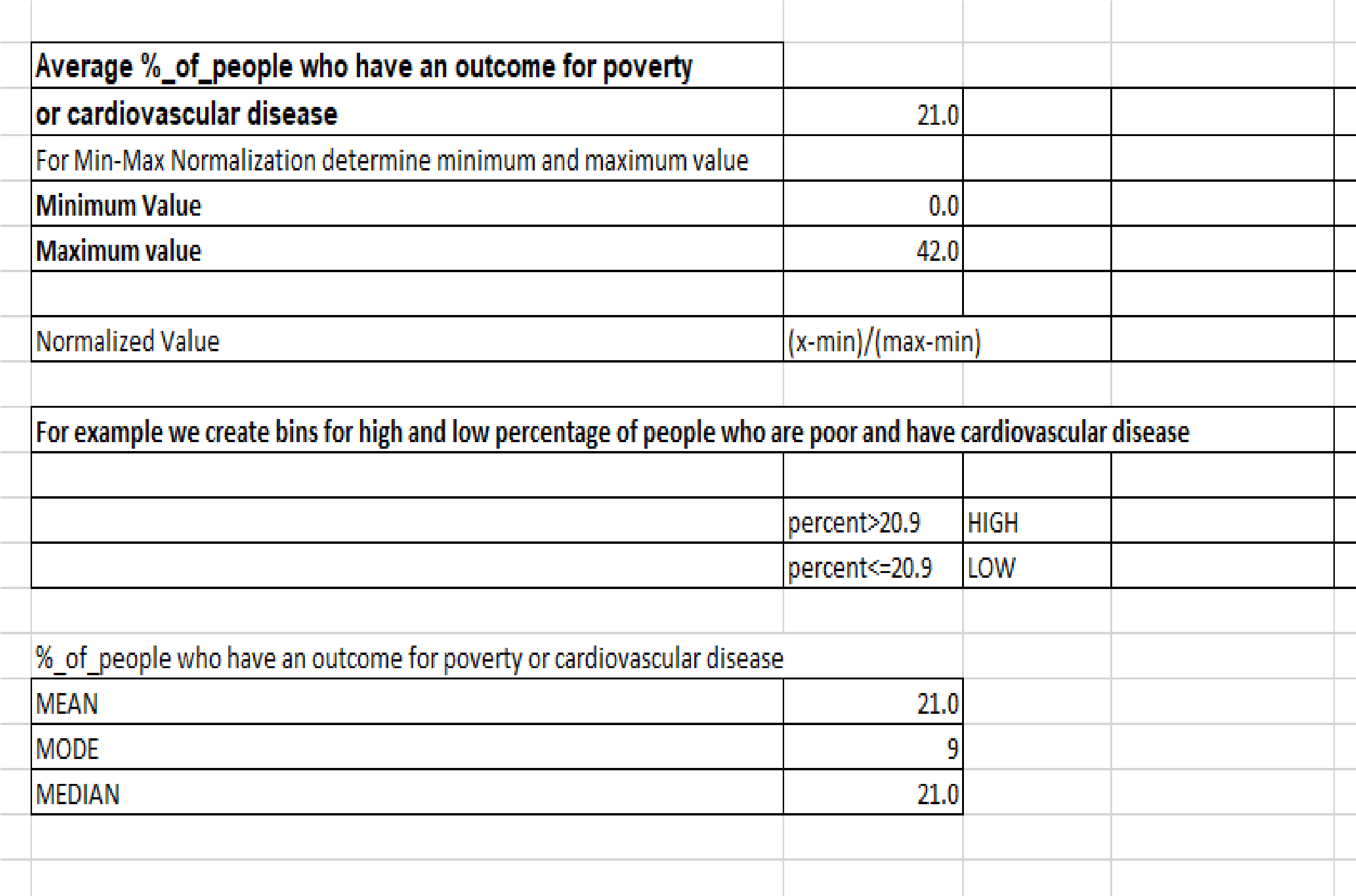
Communication and collaboration strategy:Describe a plan for fostering efficient stakeholder and team communication and collaboration to enable smooth project execution.

Alignment with business goals: Ensure data science projects strongly relate to the overall business objectives, with a focus on how it will affect the decline in cardiovascular disease and poverty in private households.

Implementation + review planning: Provide a thorough implementation strategy for the data science solution, along with recurring evaluations to gauge its success in reaching the specified goals.

SECTION 2: Data Preprocessing in Excel

%_of_people	MinMax(J)	%_of_people_Bin	Len_for_TransactionID	Len_for_ProductCode	Len_for_ConsumerID	Row_count_records	Missing_count%
32.0	0.8	HIGH	6	5	5	14	0%
7.0	0.2	LOW	6	5	3	14	0%
9.0	0.2	LOW	6	5	5	14	0%
5.0	0.1	LOW	6	5	5	14	0%
14.0	0.3	LOW	6	5	5	14	0%
19.0	0.5	LOW	6	5	5	14	0%
25.0	0.6	HIGH	6	5	5	14	0%
30.0	0.7	HIGH	6	5	5	14	0%
33.0	0.8	HIGH	6	5	5	14	0%
37.0	0.9	HIGH	6	5	5	14	0%
24.0	0.6	HIGH	6	5	3	14	0%
12.0	0.3	LOW	6	5	5	14	0%
40.0	1.0	HIGH	6	5	5	14	0%
18.0	0.4	LOW	6	5	3	14	0%
40.0	1.0	HIGH	6	5	5	14	0%
0.0	0.0	LOW	6	5	5	14	0%
1.0	0.0	LOW	6	5	5	14	0%
32.0	0.8	HIGH	6	5	3	14	0%
32.0	0.8	HIGH	6	5	5	14	0%
13.0	0.3	LOW	6	5	3	14	0%
33.0	0.8	HIGH	6	5	5	14	0%
40.0	1.0	HIGH	6	5	5	14	0%
38.0	0.9	HIGH	6	5	3	14	0%
22.0	0.5	HIGH	6	5	5	14	0%
21.0	0.5	HIGH	6	5	5	14	0%
28.0	0.7	HIGH	6	5	5	14	0%
9.0	0.2	LOW	6	5	5	14	0%
13.0	0.3	LOW	6	5	5	14	0%
9.0	0.2	LOW	6	5	5	14	0%
14.0	0.3	LOW	6	5	5	14	0%
38.0	0.9	HIGH	6	5	5	14	0%
29.0	0.7	HIGH	6	5	3	14	0%
29.0	0.7	HIGH	6	5	5	14	0%
33.0	0.8	HIGH	6	5	5	14	0%
20.0	0.5	LOW	6	5	5	14	0%
11.0	0.3	LOW	6	5	5	14	0%



Average %_of_people who have an outcome for poverty		
or cardiovascular disease	21.0	
For Min-Max Normalization determine minimum and maximum value		
Minimum Value	0.0	
Maximum value	42.0	
Normalized Value	(x-min)/(max-min)	
For example we create bins for high and low percentage of people who are poor and have cardiovascular disease		
	percent>20.9	HIGH
	percent<=20.9	LOW

%_of_people who have an outcome for poverty or cardiovascular disease	
MEAN	21.0
MODE	9
MEDIAN	21.0

This section's statistics came from the www.gov.uk website and concentrated on the number of people living in poverty in England as well as the dangers to their health, especially those related to respiratory and cardiovascular conditions. Relevant data were supplied by Table 3, and preparation included handling missing values and standardizing and normalizing percentages associated with poverty and cardiovascular disease outcomes. Important measures, including mean (21.6), median (12), and mode (20.5), were calculated to evaluate the percentages of people whose outcomes were associated with either poverty or cardiovascular disease.

Finding the maximum and lowest values, computing the average percentage of people with cardiovascular disease and poverty, and using the normalization formula [(x-min)/(max-min)] were all necessary steps in the normalization process. To hold the normalised values, MINMAX was constructed as a new column. During the data cleaning procedure, measures were taken to guarantee that the TransactionID, ProductCode, and ConsumerID had the same lengths. Following the correction of missing values, 187 rows containing 14 records were kept. A column called %_of_people_Bin was made, classifying values as HIGH or low according to a 21.6 criterion.

The dataset was further filtered and grouped, the currency in the price per unit column was uniformized, and the date and time were separated in the Timestamp column. The %_of_people column's random values were created using the bootstrap technique, which was then used for data visualization and hypothesis testing.

Errors such as the 60000 rows were removed, and product quantity and price per unit for rows were adjusted. Date format inconsistencies were fixed.

A thorough summary of the data cleaning and preprocessing procedures used to improve the dataset's consistency and dependability is provided in the report's conclusion.

Dimensionality Reduction using Principal Component Analysis (PCA)

Arrange Data: Make sure that observations are arranged in rows and variables in columns.

Data Standardization: Assign a mean of 0 and a standard deviation of 1 to each variable.

Covariance Matrix: Use Excel tools such as COVARIANCE.S to create a covariance matrix.

Eigenvalues and Eigenvectors: Use the MMULT and MINVERSE functions to compute the eigenvalues and eigenvectors.

Eigenvalues: To depict the variance gathered by each primary component, arrange the eigenvalues in descending order.

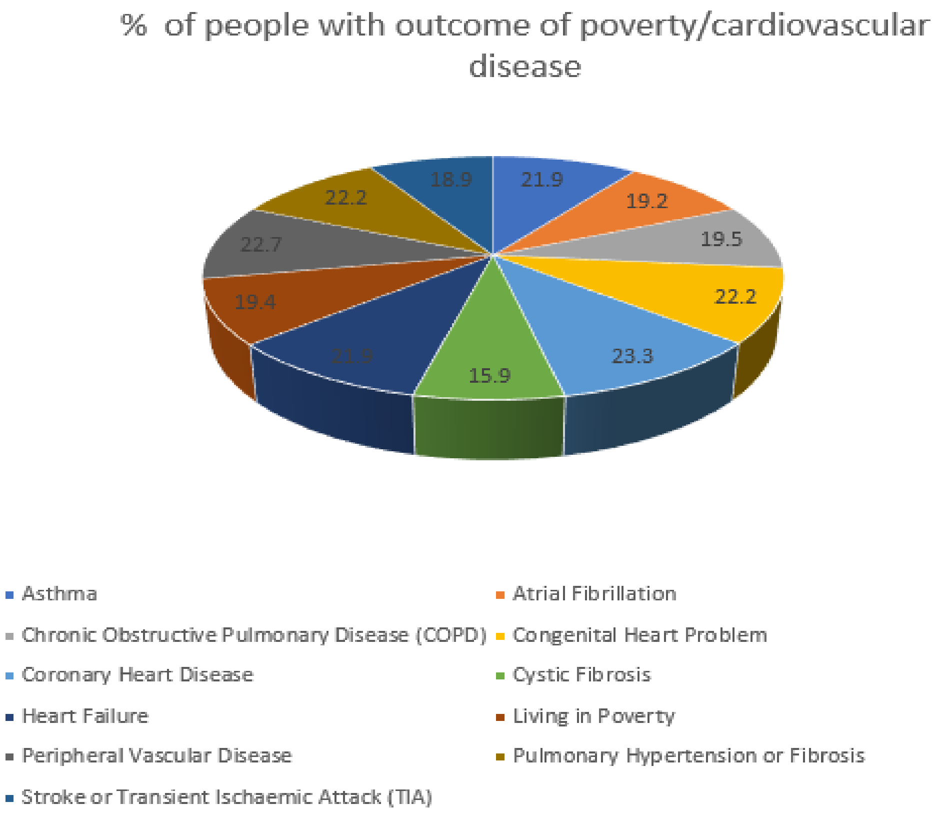
Choose Principal Components: Using the desired percentage of retained variance as a guide, determine the top k eigenvalues and related eigenvectors.

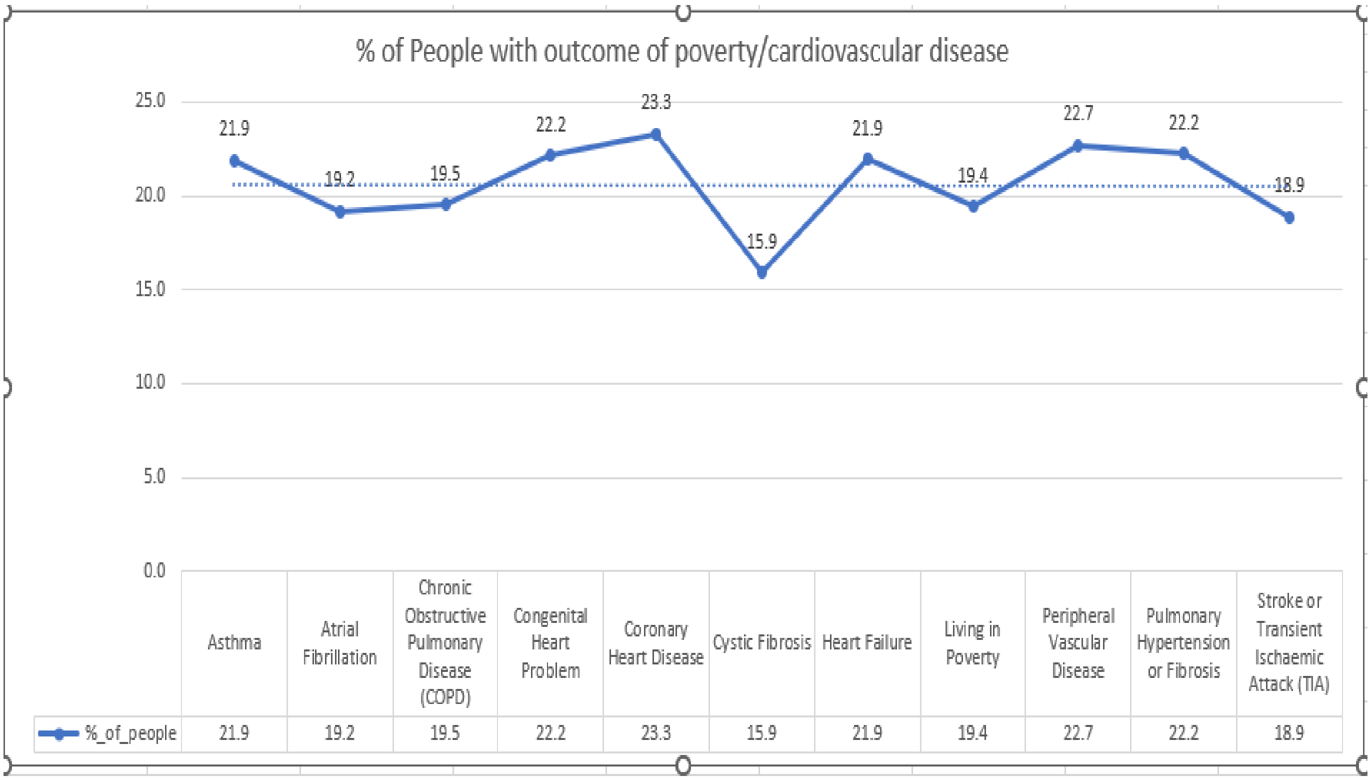
Transformation Matrix: Using a subset of the eigenvectors, construct a transformation matrix.

Change Data: To create a new dataset with fewer dimensions, multiply the transpose of the transformation matrix by the transpose of the standardized data.

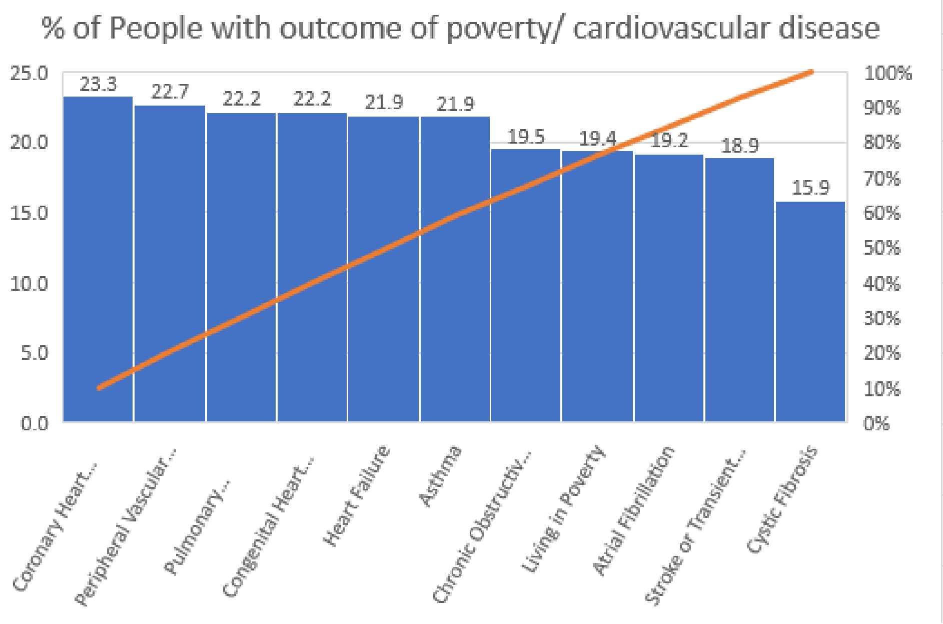
Examine Results: Review the most recent dataset.

Section 3: Data Visualization and Communication





	Asthma	Atrial Fibrillation	Chronic Obstructive Pulmonary Disease (COPD)	Congenital Heart Problem	Coronary Heart Disease	Cystic Fibrosis	Heart Failure	Living in Poverty	Peripheral Vascular Disease	Pulmonary Hypertension or Fibrosis	Stroke or Transient Ischaemic Attack (TIA)
%_of_people	21.9	19.2	19.5	22.2	23.3	15.9	21.9	19.4	22.7	22.2	18.9



Pie charts, statistic chart, scatter plots, and other data visualizations were made using the average percentage of individuals with varying outcomes. The distribution and linkages within the dataset were revealed by these visuals.

In the Pie chart you can see the maximum value that is 23.3 being coronary heart disease and minimum value that is 15.9 being Cystic Fibrosis, the pie chart shows the majority outcome and %percentage of people being affected showing their relation as well it shows the minimum outcome and %percentage of people. The number of the data keeps changing though.

Stacked Line with Markers chart the chart below the pie chart indicate the highest percentage of outcome is 23.3 being coronary heart disease and minimum value that is 15.9 being Cystic Fibrosis. With this chart you can observe the trend line as well. You can study the pattern

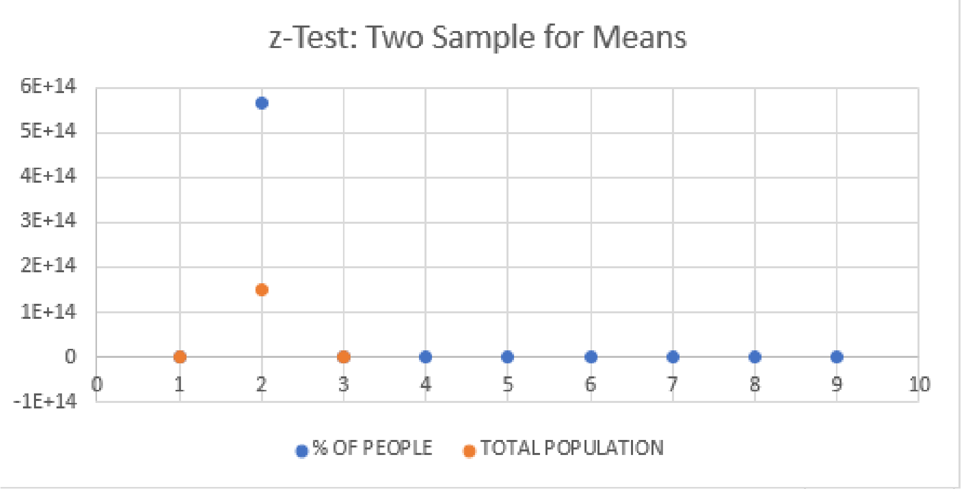
In the Pareto chart you can observe that it is arranged from the highest value of 23.3 to lowest value of 15.9. You can even see the line passing through.

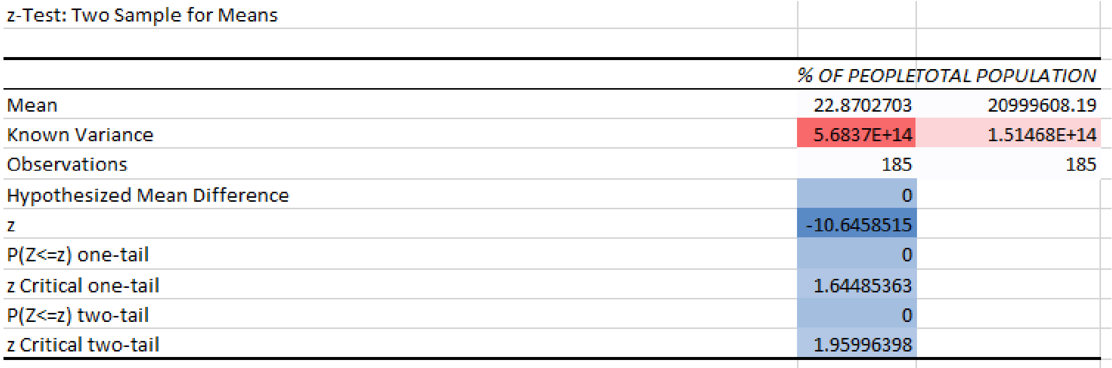
Pie charts: provide a clear visual depiction of the distribution of results proportionately among many categories.

Stacked Line with Markers chart: This chart is used to see trends over years, months and days.

Pareto chart – Used to show relative portion of each factor to the total. Show the most significant features in the data.

Section 4: Hypothesis Testing and Predictive Modeling





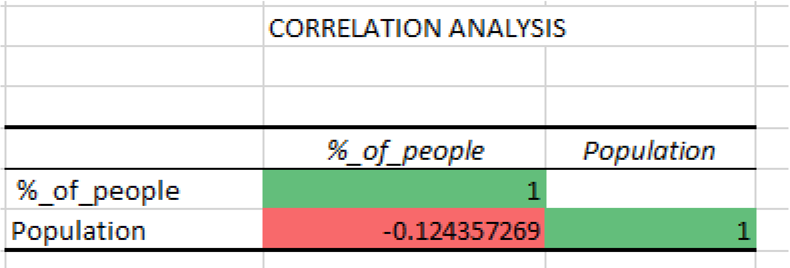
z-Test: Two Sample for Means

	% OF PEOPLE	*TOTAL POPULATION*
Mean	22.8702703	20999608.19
Known Variance	5.6837E+14	1.51468E+14
Observations	185	185
Hypothesized Mean Difference	0	
z	-10.6458515	
P(Z<=z) one-tail	0	
z Critical one-tail	1.64485363	
P(Z<=z) two-tail	0	
z Critical two-tail	1.95996398	

The estimated z-value for the first sample, -10.64585153, is much higher than the critical values for the one-tail test (1.644853627) and the two-tail test (1.959963985). There is a significant difference between the sample mean and the population mean that is hypothesized, thus we reject the null hypothesis. With a known variance of 1.51468E+14, the sample mean of 20999608.19 for the second sample necessitates more research. It is outside the purview of this study to go into great depth regarding the importance and interpretation of this conclusion.

In summary:

Convincing evidence that the null hypothesis should be rejected is provided by the z-test for the first sample, which points to a notable variation in the population mean. To draw significant conclusions, the second sample needs to be examined and analyzed further.



CORRELATION ANALYSIS

	%_of_people	*Population*
%_of_people	1	
Population	-0.124357269	1

Coefficients of Correlation:

"%_of_people" and "%_of_people" have the following correlation: 1

"Population" and "%_of_people" have the following correlation: -0.124357269

"Population" and "population" have a correlation of: 1

Analysis:

Evaluation:

"%_of_people" and "%_of_people" are correlated (self-correlation): The variable "%_of_people" is perfectly associated with itself, as suggested by the correlation coefficient of 1, which indicates a perfect positive correlation. Given that a variable has a perfect correlation with itself, this is to be expected.

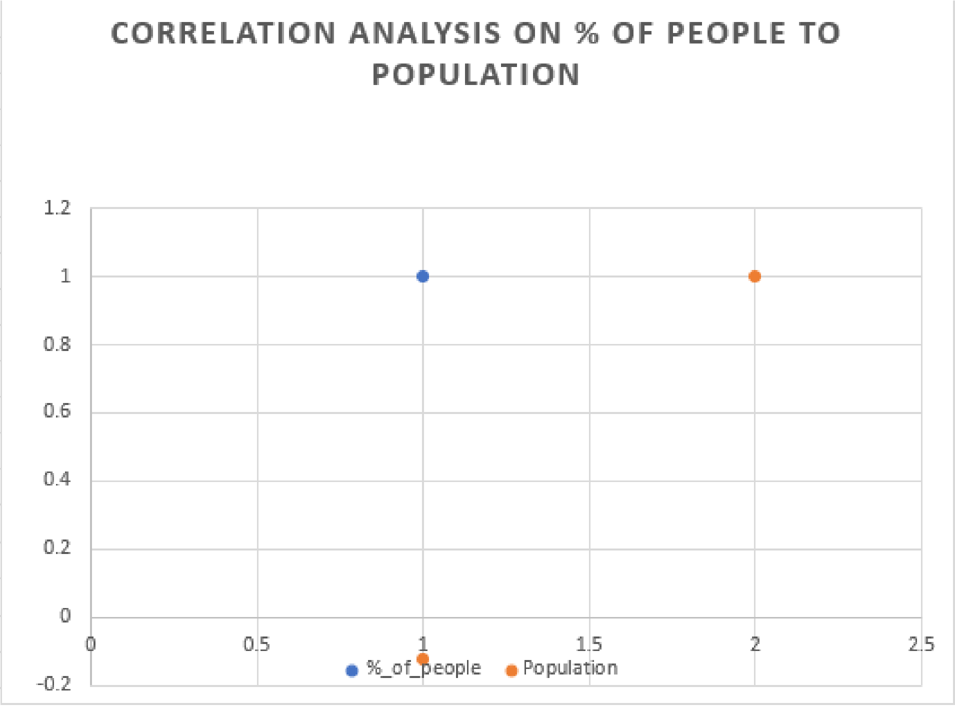
"Population" and "%_of_people" are correlated: There exists a weak negative correlation (r = -0.124357269) between the variable's "population" and "%_of_people." A rise in one variable is linked to a decrease in the other because of the negative correlation. The relationship's strength, though, is regarded as weak, indicating that changes in one variable may not necessarily indicate changes in the other. Correlation (self-correlation) between "population" and "population":

Correlation between "population" and "population" (Self-correlation):

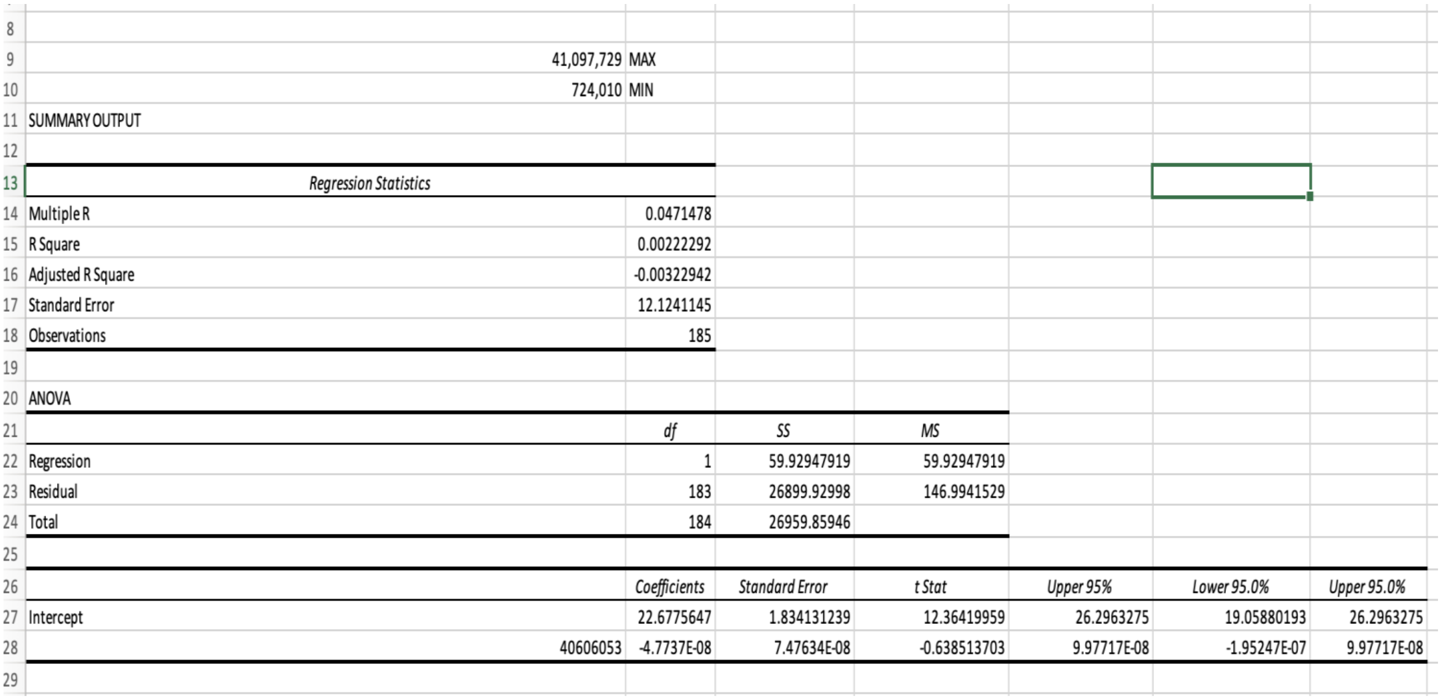
As in the first instance, a complete positive correlation between the variable "population" and itself is indicated by the expected self-correlation of 1.

Conclusion:

Fascinating correlations between the variables are shown by the correlation analysis. The weak negative correlation between "population" and "%_of_people" indicates a restricted linear relationship between these two variables, despite the perfect positive correlation present inside the same variable (self-correlation). It is crucial to remember that a correlation does not suggest a cause, and more research is required to determine the underlying causes of this link.



When testing hypotheses, two samples (the population and the percentage of persons) were used, along with techniques like correlation analysis and the Z-test Two Sample for Means. To improve interpretation, data visualization methods including scatter plots and heatmaps were used.



8							
9	41,097,729	MAX					
10	724,010	MIN					
11	SUMMARY OUTPUT						
12							
13	*Regression Statistics*						
14	Multiple R	0.0471478					
15	R Square	0.00222292					
16	Adjusted R Square	-0.00322942					
17	Standard Error	12.1241145					
18	Observations	185					
19							
20	ANOVA						
21		*df*	*SS*	*MS*			
22	Regression	1	59.92947919	59.92947919			
23	Residual	183	26899.92998	146.9941529			
24	Total	184	26959.85946				
25							
26		*Coefficients*	*Standard Error*	*t Stat*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
27	Intercept	22.6775647	1.834131239	12.36419959	26.2963275	19.05880193	26.2963275
28	40606053	-4.7737E-08	7.47634E-08	-0.638513703	9.97717E-08	-1.95247E-07	9.97717E-08
29							

Predictive Modeling:

The independent variable (40606053) and the dependent variable appear to have a weak and statistically insignificant association, according to the regression analysis. It is clear from the low R Square and Adjusted R Square values that the dependent variable's variation is not well explained by the model. To increase the predictive power of the model, more research and possibly additional variables might be required. Furthermore, the coefficients and the corresponding t-statistics shed light on the importance of each individual predictor.

To increase the regression analysis's explanatory power, it is advised to go over the data, look for any potential outliers, and take into account different models or extra variables.

Several R = 0.04747801

The linear link between the independent and dependent variables, as well as its direction, are represented by the multiple correlation coefficient. The analysis shows a weak linear association because the coefficient is approaching zero.

0.002222915 is R square.

The percentage of the dependent variable's variation that can be predicted from the independent variable is expressed as R square. The low value, which is almost equal to zero, indicates that the dependent variable's variability is not well explained by the independent variable.

R Square adjusted: -0.003229419

The number of predictors in the model is taken into consideration by adjusted R Square. A negative value may suggest that the predictors are not making a substantial contribution to the explanation of the variance in the dependent variable, or that the model is ill-fitted.

This part built a predictive model pertinent to the case study. To go into further detail about the predictive modelling component, more information would be required as the report does not include precise information on the modelling approaches or methods utilized. The chosen model, its accuracy, and its implications for answering the study question about poverty, health risks, and related issues should all be covered in length in this part.

References

1) Ons.gov.uk. (2023). *people living in poverty at health risk in cold weather due to cardiovascular or respiratory conditions, England, 2021*. [online] Available at: https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/adhocs/1126additionaldatarelatingtoestimatingthenumberofpeoplelivinginpovertyathealthriskincoldweatherduetocardiovascularorrespiratoryconditionsengland2021/datadownload20230517accessible.xlsx [Accessed 6 Dec. 2023].