Research Study for Scientia Group

Research on the relevancy of using car plates numbers to assess a group socio-economic status.

Date: 05/11/2023

Table of Content

Background and motivation

Understanding socio-economic status (SES) within urban communities is crucial for policymakers, economists, and social scientists alike, as it impacts decision-making processes from urban planning to welfare distribution. Traditionally, SES has been measured using census data, which, while comprehensive, often lag behind the rapidly changing urban landscapes. Hence, there is a pressing need for more dynamic, real-time indicators of SES that can reflect the current status of a community. The Scientia study group, inspired by the need for innovative methods to gauge SES, identified car ownership as a potential indicator. Car ownership patterns may provide immediate insights into the economic wellbeing and lifestyle choices of residents, as the type and age of a vehicle can be indicative of financial status and spending priorities. This premise is grounded in the theory of conspicuous consumption and has practical implications for understanding SES outside of traditional census metrics.

The motivation for this research emerges from two critical observations: firstly, the dynamic nature of car ownership in response to economic shifts, and secondly, the practicality of collecting and analysing vehicle registration data as a proxy for wealth and economic status. By examining the correlation between the age of cars and SES within different suburbs, the group aims to establish a novel, timely, and possibly more reactive measure of socio-economic conditions.

Moreover, this approach has the potential to uncover patterns and socio-economic insights at a granular level, which may not be evident through standard census data. For instance, car ownership details can provide immediate information post-economic events, such as a recession or a boom, and thus offer a leading indicator to the changing SES of a community. Such data can be invaluable for government agencies and non-governmental organizations in crafting timely and responsive social policies.

To actualize this concept, the data collection process commenced on 8th August 2023 and was completed by 4th October 2023. A total of 30 team members dedicated themselves to gathering data across 6 suburbs in South Australia, resulting in an initial collection of 7,908 license plates. This extensive collection effort sets the foundation for a comprehensive analysis that seeks to illuminate the correlation between car ownership and socio-economic status in a concrete and measurable way.

Data Explanation

Data Collection

The data collection process began on 8th August 2023 and was completed by 4th October 2023. A total of 30 team members participated in gathering the data across various suburbs in South Australia. The primary objective was to collect license plate data and the suburb they were collected from, which resulted in an initial collection of 7,908 license plates.

Data Cleaning Process: The raw data underwent several cleaning steps to ensure its accuracy and relevance:

- **Standardization**: All plate numbers were standardized by removing hyphens ("-") and spaces to ensure consistent formatting.
- **Invalid Plates Removal**: License plates containing asterisks (*) were removed.
- **Length-based Filtering**: Plates with lengths between 1-5 or more than 7 characters were excluded.
- **Unique Plates Extraction**: Duplicate plates were identified and removed, retaining only unique entries.
- **Missing Values**: Columns with missing values and no relevant data were removed for data streamlining. This includes car plates with missing postcode.

Data Enrichment: Post-cleaning, the dataset underwent enrichment:

- **Car Data Acquisition**: Using the Australian CAR Report's web API, the unique license plates were matched to retrieve the car's manufacturing year and its estimated value. Car plates which did not returned a manufacturing year using the API where removed from the dataset. Website link: https://aucn.net.au/australian-car/test/Home/?vin=S694%20BSB&state=SA&frm=aucar
- **SES Index Addition**: The "Score of Index of Relative Socio-economic Disadvantage" (SES index) was incorporated into the database from the SEIFA 2021 document. This index provides insights into the socio-economic status of the areas linked with the postcodes. This addition aims to explore the relationship between car age and socio-economic status in South Australia. Website link: https://www.abs.gov.au/statistics/detailed-methodology-information/concepts-sources-methods/socioeconomic-indexes-areas-seifa-technical-paper/2021/construction-indexes#geographic-output-levels-forseifa-2021

Final Dataset Structure: The enriched dataset comprises:

- **Car plates**: License plate numbers.
- **Postcode**: Postal codes tied to the license plates.
- **Year**: Year of car manufacture.
- **Age:** Year – Current year
- **Average Dealer Value**: Estimated vehicle values via the Australian CAR Report's web.
- **SES Index**: Socio-economic disadvantage score derived from the SEIFA 2021 data.

Data Quality

Timeliness

Timeliness refers to the extent to which the data accurately represent the reality of car number plates at the specific point in time when they were collected. Each field in the collected data set can be considered to discuss how each of them corresponds to timeliness.

Completeness

Completeness is achieved by capturing all types of car number plates in the suburbs, accounting for variations like different car number plate types (standard, personalised, specialty), and ensuring accuracy of the collected data. Each of these fields can be examined in the context of the concept of completeness during a data quality assessment.

Uniqueness

Uniqueness in the context of the car number plate dataset signifies that no instance of a specific car number plate is recorded more than once within the dataset. The relationship of each field to the concept of uniqueness during a data quality assessment can be explored.

Validity

Validity, within the context of the car number plate dataset, implies the degree to which the data adheres to predetermined rules, standards, or specifications about the structure, category, and permissible values linked to car number plates. Each field concerning the concept of validity during a data quality assessment can be assessed.

Accuracy

Data accuracy for the car number plate dataset refers to the extent to which the recorded number plates correctly represent the vehicles in the surveyed suburbs. It assesses the precision of the dataset in comparison to the actual number plates of the vehicles captured during the survey. Here's a breakdown of how each field relates to accuracy.

Consistency

Consistency in a data quality assessment refers to the uniformity and reliability of data across the dataset. Data should be captured, represented, and processed in a consistent manner. Here's an examination of each field in relation to consistency.

Further details on data quality assessment and selection bias are discussed in Appendix 02 – Data Quality Assessment.

Analytic Approach

Data Exploration and Familiarization

Dataset Loading: Uploaded the dataset into the analytical environment for integrity.

Structural Analysis: Examined the dataset structure to confirm variable integrity and formatting.

Descriptive Statistics: Generated statistics for central tendency, dispersion, and distribution.

Data Segmentation: Segmented the dataset by postcodes for detailed socio-economic analysis.

Median Age Calculation

Aggregation by Suburb: Calculated the median car age for each suburb to minimize outlier effects.

SES Score Retrieval: Extracted SES scores from SEIFA 2021 for accuracy.

Data Merging for Comparison

Data Synthesis: Merged car age data with SES scores based on postcodes.

Comparative Analysis: Created a dataset for direct comparison of median car ages to SES scores.

Correlation Analysis

Correlation Hypothesis: Hypothesized a significant correlation between car age and SES scores.

Coefficient Calculation: Calculated the Pearson correlation coefficient for the linear relationship.

Statistical Significance: Tested the significance of the correlation to rule out random correlations.

Visualization for Insight Generation

Graphical Representation: Constructed combined bar and line diagrams for median car age and SES scores.

Interpretative Visualization: Designed visualizations for pattern identification and outlier analysis.

Visualization Refinement: Iterated on design for clarity and alignment with the data narrative.

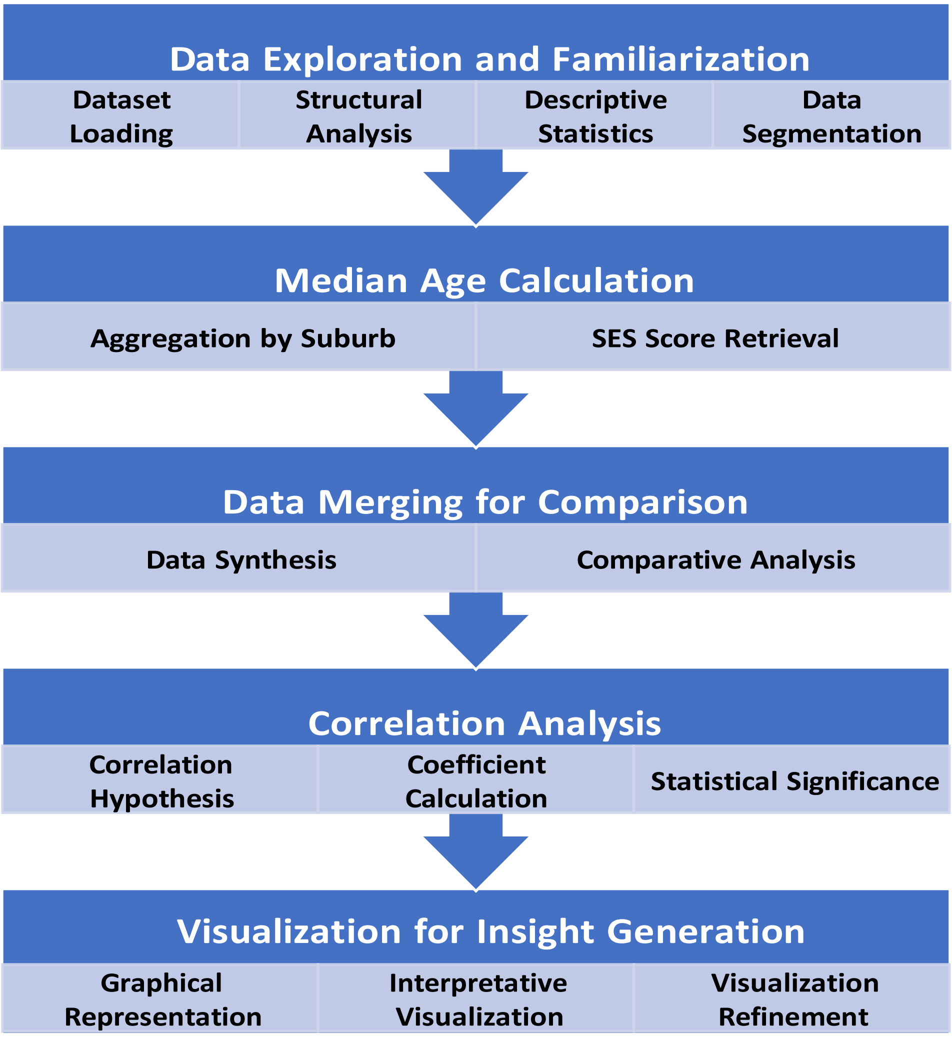

Figure 01: Schematic Diagram for the Analytical Approach

Table of Median Age of Cars and SES Score by Postcode

Postcode	**Median Age of Cars**	**SES Score**
5007	10 years	976.52
5031	10 years	990.85
5061	11 years	1097.26
5081	9 years	1079.86
5084	15 years	863.33
5085	13 years	1034.93
5086	10 years	995.20
5094	14 years	1039.88
5095	11 years	1034.93
5107	10 years	880.40
5112	9 years	717.34
5251	8.5 years	1018.86

Table 01: Table showing postcode, median age of cars and Score of Index of Relative Socio-economic Disadvantage.

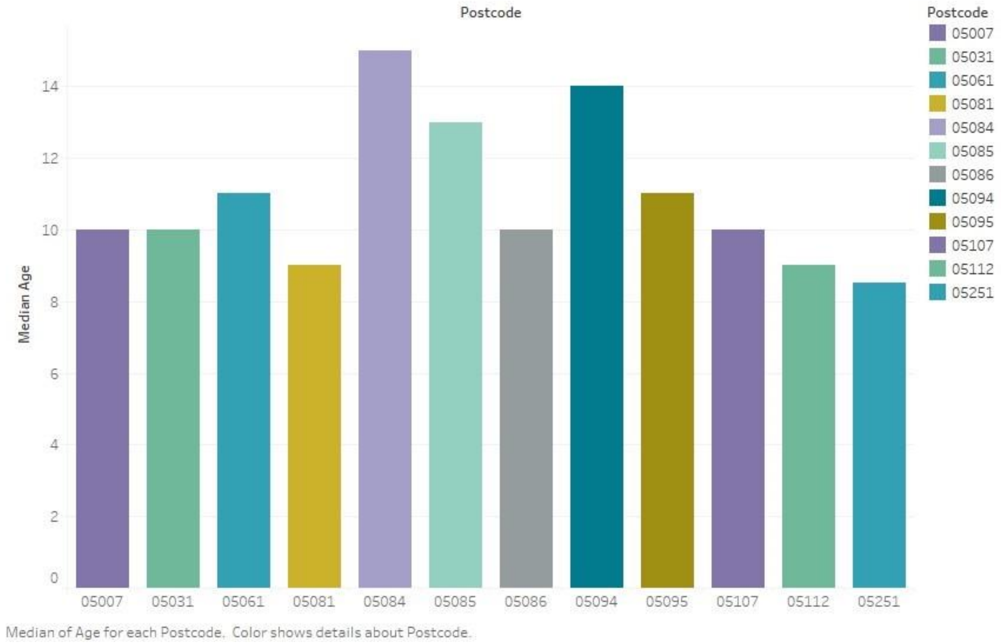

Figure 02: Graph showing postcode vs median age of cars

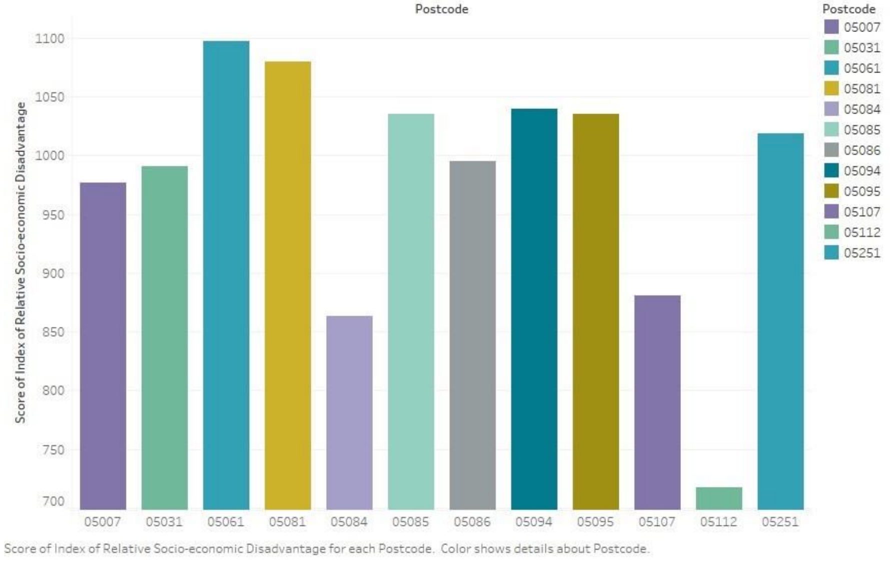

Figure 03 : Graph showing postcode vs SES values

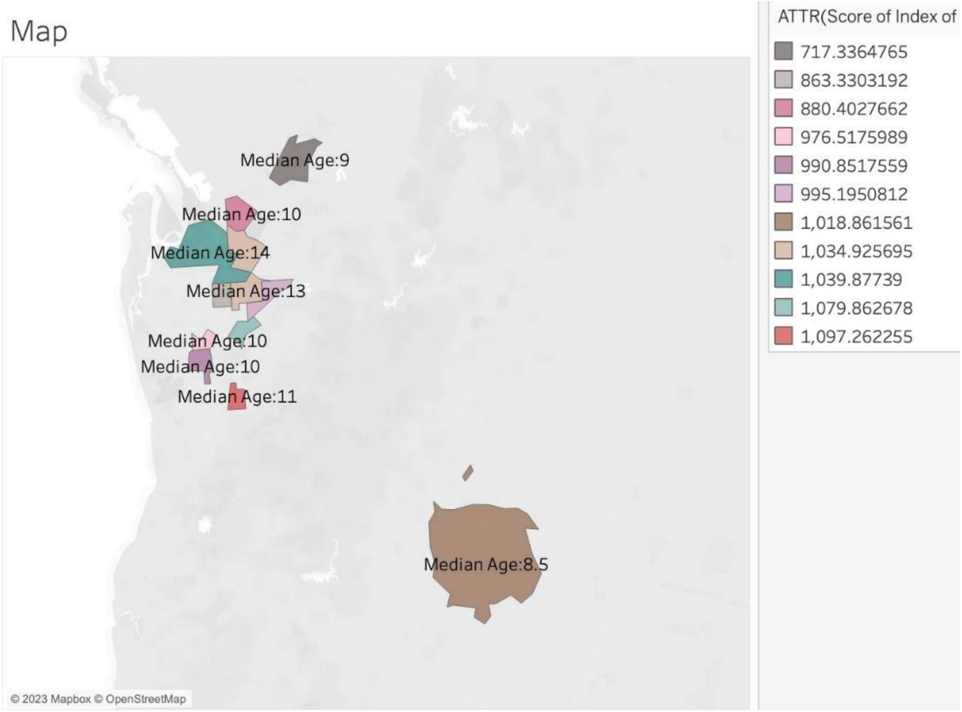

Figure 04: Map showing SES Score and Median Age

From this table

- Suburb 5061 has a relatively older median car age (11 years) and the highest SES score (1097.26).
- Suburb 5112 has the youngest median car age (9 years) but the lowest SES score (717.34).
- Suburb 5251 has a young median car age (8.5 years) with a decent SES score of 1018.86.

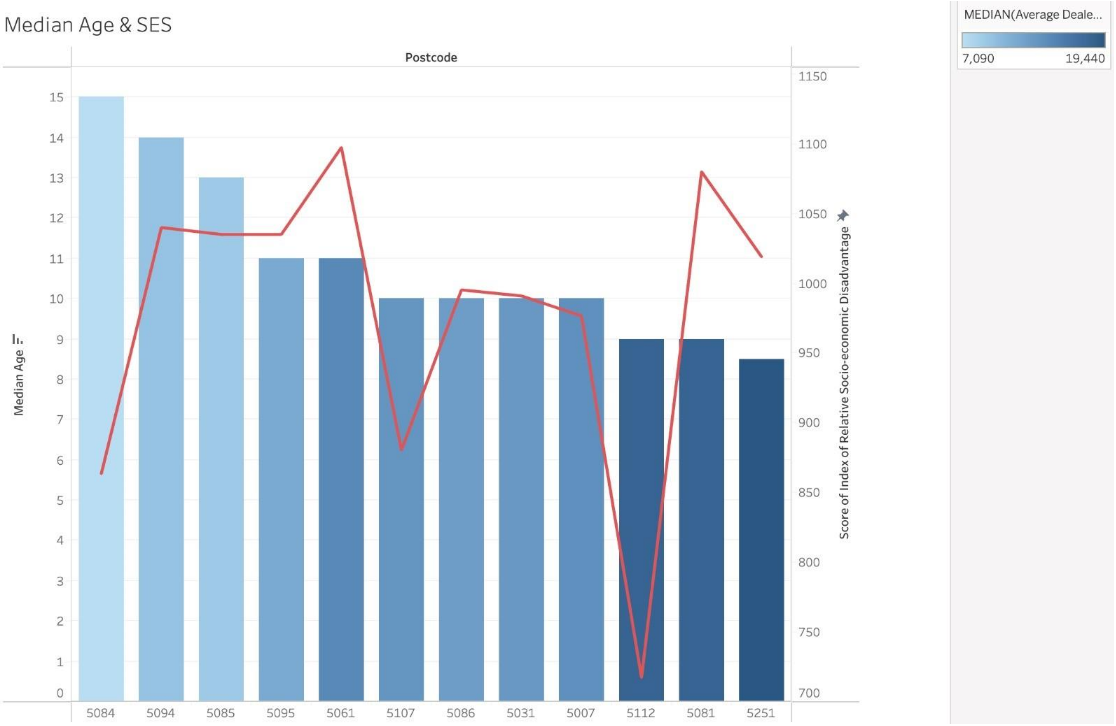

Figure 05: Comparing the median age of cars and the SES score for each suburb

This diagram comparing the median age of cars and the SES score for each suburb:

- The blue bars represent the median age of cars in each suburb, the shade of blue represents the value of the car; the darker the color, the more valuable the car is.
- The red line with markers represents the SES score for each suburb.
- Suburb with postcode 5061 has both a relatively older median car age (11 years) and the highest SES (1097.26).
- Suburb 5112, on the other hand, has a median car age of 9 years but the lowest SES (717.34).
- Suburb 5251 has the youngest median car age (8.5 years) but a decent SES score (1018.86).

Pearson Correlation Coefficient Calculation

The formula for the Pearson correlation coefficient (r) for two variables X and Y is:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

- n is the number of paired data points.

- Σxy is the sum of the product of each pair of values.

- Σx and Σy are the sums of the X values and Y values, respectively.

- Σx^2 and Σy^2 are the sums of the squares of X values and Y values, respectively.

Using the provided data:

- X represents "Median Age of Cars".

- Y represents "SES Score".

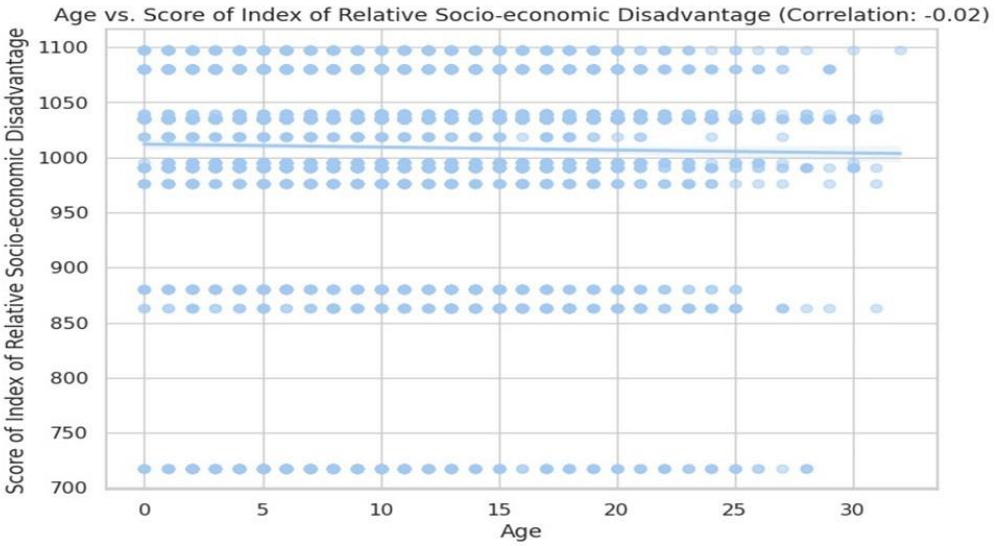

Figure 06: Age Vs SES value Scatter plot

From our calculations, the Pearson correlation coefficient between "Median Age of Cars" and "SES

Score" is approximately r ≈-0.0212

This means that there's a very weak positive linear relationship between the two variables. Specifically, as the median age of cars increases, the SES score might slightly increase, but this relationship is so weak it's almost non-existent.

Key Insights:

- There's no strong linear correlation between the age of cars and the socio-economic status based on this dataset.
- The visualization provided insights into suburbs with relatively young car ages and their corresponding SES scores, showing that socio-economic status doesn't necessarily dictate car age in all suburbs.

Findings

1. Interpreting the Correlation Value: The Pearson correlation coefficient between "Median Age of Cars" and "SES Score" is approximately r ≈-0.0212. This means that there's very little linear association between "Median Age of Cars" and "SES Score". In practical terms, knowing the value of one variable gives us very little information about the expected value of the other variable.

2. Interpreting the Relationship: The negative sign in the correlation value suggests that there's a slight inverse relationship between the two variables. If there was any noticeable trend, it would imply that as the "Median Age of Cars" goes up, the "SES Score" might go down, and vice versa. However, given the value's closeness to zero, this trend is negligible.

3. The Magnitude of the Relationship: The magnitude (or absolute value) of r tells us about the strength of the relationship. As mentioned, the closer r is to 0, the weaker the relationship. An r value of -0.0212, in absolute terms, indicates an extremely weak relationship. This means that fluctuations in one variable are not meaningfully reflected in changes in the other variable.

4. Practical Solution: In real-world scenarios, such a weak correlation might suggest that other factors not considered in this analysis could be driving the "SES Score". It would be unwise to make policy or business decisions based solely on this relationship, as the two variables don't seem to influence each other in any significant way.

5. Causation vs. Correlation: It's vital to note that correlation does not imply causation. Even if there was a stronger correlation, it wouldn't necessarily mean that changes in the "Median Age of Cars" cause changes in the "SES Score", or vice versa. There could be other lurking variables or factors at play.

6. Conclusion: To sum up, the analysis suggests that there's almost no linear relationship between the "Median Age of Cars" and the "SES Score". For stakeholders or decision-makers, this means they should look elsewhere for factors or variables that might more significantly impact or explain the variation in "SES Score".

Data Interpretation

The endeavour to understand the dynamics between car ownership patterns and socio-economic standings led to a comprehensive data collection and analysis exercise spanning various suburbs in South Australia. The primary aim was to ascertain whether there exists a linear relationship between the age of cars and the Socio-economic Disadvantage Score (SES) derived from the SEIFA 2021 data. Utilising a robust dataset, constructed post meticulous data cleaning and enrichment procedures, a thorough analysis was performed to gauge the correlation between the median age of cars and the SES score across different suburbs. The Pearson correlation coefficient was employed as the key statistical measure to evaluate the linear association between the two variables.

The resultant correlation coefficient of approximately $r \approx -0.0212$ elucidates a very weak, almost nonexistent linear relationship between car age and SES score. This inconclusive correlation prompts a deeper examination of various factors that might be influencing car age and SES score independently or interactively. Through a detailed exploration, it has been discerned that multiple external and intrinsic factors, ranging from cultural norms to economic conditions, could be playing pivotal roles in shaping car ownership behaviours and socio-economic landscapes across the suburbs, thereby diluting the direct linear relationship between car age and socio-economic status. This analysis serves as a steppingstone towards a more nuanced understanding of the multifaceted interactions between socio-economic indicators and car ownership patterns, providing a foundation for further research in this domain.

1. Data Bias:

Variability within Suburbs:

The variability in car ages within each suburb could be high, leading to a weak correlation when aggregated at the suburb level. A suburb with a mix of old and new cars could have a median car age that doesn't reflect the socioeconomic status of its residents. For instance, a suburb might have a wide range of residents from different socioeconomic backgrounds. Some wealthier residents might be able to afford to buy new cars frequently, hence having newer cars, while less affluent residents might hold onto their cars for longer, resulting in older cars. When aggregating the data at the suburb level to calculate the median age of cars, this diversity can lead to a median value that does not accurately reflect the socioeconomic status of the entire suburb. The median age of cars might fall somewhere in the middle of the range of car ages, which could mask the true relationship between car age and socioeconomic status. For example, if a suburb has a good mix of very old (e.g., 15+ years) and very new cars (e.g., 1-2 years), the median age might be around 8 years, which doesn't provide a clear insight into the socioeconomic standing of the suburb's residents. This scenario could potentially dilute any existing correlation between the age of cars and the socioeconomic disadvantage score derived from the SEIFA 2021 data, thereby leading to a near-zero correlation coefficient as observed in the analysis.

Other Influencing Factors:

There might be other factors influencing car age and SES score that were not accounted for in this analysis. For instance, personal preferences might play a significant role in determining car age; some individuals may prefer newer models due to their modern features and aesthetics, while others might prefer older models for their classic design or lower cost. Similarly, family size could influence car age, as larger families might require bigger or newer cars to accommodate everyone comfortably and safely. Moreover, the availability and accessibility of other forms of transportation like public transit, cycling, or walking paths in a suburb could impact car ownership patterns. For example, in suburbs with robust public transportation systems, residents might not feel the need to update their cars as frequently, or they might opt for older, less expensive models since the car may not be their primary mode of transport. On the other hand, the SES score, which is an indicator of socio-economic status, might be affected by a range of factors like employment rates, education levels, and housing conditions, which were not considered in the analysis. The interplay of these various factors could create a complex scenario where the relationship between car age and SES score is obscured or overshadowed by other influencing factors, leading to the observed weak correlation in the dataset.

Inadequate Sample Size:

The sample size or the number of suburbs analysed might not be large enough to provide a clear picture of the relationship between car age and SES score. For instance, if only a handful of suburbs were analysed, and these happened to be areas with peculiar car ownership patterns due to local circumstances (like a recently opened car dealership offering discounts on new cars), the observed relationship (or lack thereof) between car age and SES score in these specific suburbs might not reflect the true relationship across a wider array of suburbs in South Australia. Additionally, a small sample size reduces the statistical power of the analysis, making it difficult to detect significant relationships even if they exist. For example, if the analysis only covered 10 suburbs out of several hundred in South Australia, the findings might not generalise well to the entire state. Therefore, the inadequate sample size could be a significant limiting factor in the analysis, and expanding the analysis to include a larger number of suburbs might provide a clearer picture of the relationship between car age and socio-economic status across the region.

2. Cultural or Regional Factors

Cultural or regional factors might influence car ownership patterns in ways that are not captured by the SES score. For example, in some cultures or regions, owning a newer car might be seen as a status symbol, prompting individuals to update their vehicles frequently. Conversely, in other areas, there might be a strong culture of maintaining and keeping vehicles for a long period, valuing longevity and sustainability over the novelty of owning a newer model. Furthermore, regional factors such as the availability of financing options for purchasing cars, or local policies and incentives around vehicle ownership could also play a significant role. Moreover, regions with harsher climates or rough terrains might necessitate owning newer or better-maintained vehicles, influencing the age of cars in those areas. Such cultural and regional factors can create variances in car ownership patterns across different suburbs that are independent of the socio-economic conditions captured by the SES score. Therefore, these unaccounted cultural and regional nuances might contribute to the observed weak correlation between car age and SES score, as they add layers of complexity to the relationship between these variables.

3. External Economic Factors

Economic conditions such as recessions or booms can influence both car purchasing behaviours and socio-economic statuses, which might further complicate the relationship between car age and SES score. For instance, during economic booms, individuals might have more disposable income, which could lead to increased purchases of new cars, thereby lowering the median car age in a suburb. Conversely, during economic recessions, individuals might postpone buying new cars or opt for used cars to save money, leading to an increase in the median car age. The SES score, reflecting the socioeconomic status of a suburb, could also fluctuate with changing economic conditions, as employment rates, income levels, and other economic factors shift.

Recommendation for Further Research

Considering the above findings, it indicates a negligible linear relationship between the median age of cars and socio-economic status (SES) scores. Hence, it is clear that more sophisticated research is required to understand the complexities of this relationship. Below are comprehensive recommendations for future research regarding this relationship and these factors are the recommended focus of a subsequent research project which aims to provide a deeper understanding of the factors influencing car ownership and SES.

- **Adopting Advanced Analytical Techniques**

Due to the complexity of socio-economic phenomena, sophisticated statistical methods such as factor analysis, multivariate regression, and structural equation modeling can be used to point out the relationships between multiple variables. Advanced analytics techniques will also help to identify patterns within the data that a simple correlation coefficient might miss.

- **Engage Stakeholders**

It is imperative that key stakeholders, such as car owners, dealerships, policymakers, and social workers are involved. The qualitative data provided will provide insights into the factors influencing individuals' decisions regarding car ownership across various socio-economic backgrounds. These understandings can be extremely useful when quantitative data provides a partial view of the situation.

- **Prioritise Data Quality**

The research process relies on the availability of high-quality, trustworthy data. It is crucial to obtain data from reputable agencies, ensuring it is current, comprehensive, and relevant. Methods of gathering data should be transparent and consistent.

- **Effect of New Technologies**

The automotive industry is undergoing significant changes due to electric vehicles (EVs), autonomous driving, and digitisation, requiring research on how these technologies impact car ownership patterns and socioeconomic status.

- **Focus on Geographical Disparities**

Geolocation may have a major impact on the relationship between SES and car age. Researchers should compare diverse areas, including urban and rural areas, regions with varying wealth levels and public service access to identify location-based inequalities.

- **Expansion of the Scope of Variables**

To obtain a clearer picture, the research should incorporate a wider range of variables. This includes economic factors such as employment status, household income, and educational levels, as well as variables such as access to access to public transportation, urban or rural settings, and cultural attitudes towards car ownership. It is vital to capture the variability of nature of socio-economic status, which may not be sufficiently represented by a single SES score.

- **Using a Multidisciplinary Approach**

Socio-economic status and car ownership are affected by cultural trends, economic policies, technological advancements, and environmental factors. Hence, future research therefore adopts a multifaceted approach, bringing in understanding from various fields such as sociology, economics, environmental science, and urban planning. This will assist in gaining insights on how different factors interact to shape the landscape of car ownership across various socio-economic groups.

- **Policy Analysis**

It can be recommended that any future studies relating to this can also include a comprehensive policy analysis component. Therefore, it can be assessed how various policies, such as taxation policies on new vehicle purchases or incentives for electric vehicles can have an impact to the car age demographics across SES groups.

In conclusion, the study that was conducted showed a weak correlation between car median age and SES scores, indicating the need for more detailed investigation beyond simple statistical correlations. These recommendations suggest future research into the complex factors influencing socio-economic status and car ownership.

References

Census geography | Australian Bureau of Statistics. (2021, July 16). Www.abs.gov.au. https://www.abs.gov.au/census/guide-census-data/geography

Construction of the indexes | Australian Bureau of Statistics. (2023, April 27). Www.abs.gov.au. https://www.abs.gov.au/statistics/detailed-methodology-information/conceptssources-methods/socio-economic-indexes-areas-seifa-technicalpaper/2021/construction-indexes#geographic-output-levels-for-seifa-2021

Purchase report - AUCN. (n.d.). Aucn.net.au. Retrieved November 4, 2023, from https://aucn.net.au/australian-car/test/Home/?vin=S694%20BSB&state=SA&frm=aucar

Appendix 01 - Data Dictionary

The collection of all car registration plate numbers has been collected from cars parked in residential areas. The study intentionally excluded commercial zones to increase the likelihood of residents owning vehicles in the surveyed areas.

To obtain detailed information on the car registration plates that were collected, the Australian Car Network (AUCN) website (AUCN, n.d.) was used. This allowed us to access detailed vehicle information based on the car registration plate numbers. The website sources data from various agencies, including the Personal Property Security Register (PPSR), Australian Road Transport and Traffic Agency (Austroads), Australian Financial Security Authority (AFSA), National Exchange of Vehicle and Driver Information System (NEVDIS), and third-party companies, as well as AUCN's data.

Car Plates

Definition

Car Plate is a unique alphanumeric code used to identify a vehicle. Serves as a primary key for all other car-related data.

Justification

Vehicles are uniquely identified by their license plates for ownership, legal, and law enforcement purposes.

Codeset

Car Plate Numbers are divided into 3 formats,

Format 1 (Example - S429BVP, S459CWA): The string “S999XXX” is composed of the constant character S, 3 digits represented by 9, and 3 letters represented by X. This format indicates a Sequential Region Code with the "S" signifying South Australia.

Format 2 (Example - XC057Z, WUH795): The string “ZZZ999” is composed of 3 letters represented by X and 3 digits represented by 9. This format indicates a Region Sequential Code which is an older plate series.

Format 3 (e.g., KEZZ60, CC885N): The string “XX99XX” is composed of a random number of letters represented by X and digits represented by 9. This pattern appears to be more personalized or specialty plates.

Guidelines to Use

Must ensure that all car plates are consistent with the **codeset** specification outlined in the data dictionary.

Source

The collected from the cars that were physically parked in residential areas within the designated suburb.

Postcode

Definition

Postcode is a numeric code that denotes the geographical area where the car plates were collected.

Justification

Postcode connects a car's plate to a specific suburb, indicating the likely residence of the owner. It facilitates quick sorting of data.

Codeset

Postcode is a set of four-digit numbers that uniquely identifies a postal delivery area within Adelaide.

Format (Example - 5031, 5081): “9999” is composed of 4 digits represented by 9.

Postcode	**Suburb**
5095	Mawson Lakes
5031	Mile End
5081	Medindie
5007	Hindmarsh
5086	Gilles Plains
5084	Blair Athol
5251	Mount Barker
5085	Lights View
5061	Unley
5112	Elizabeth
5107	Parafield Gardens
5094	Gepps Cross

Guidelines to Use

Must ensure that the entered postcode is accurate and represents a correct geographic area.

Source

Geographical location at the time of data collection. The locations were verified using the opendatasoft website https://public.opendatasoft.com/explore/

Year

Definition

Year is a numeric code that denotes the manufactured year of the car.

Justification

The year of a car can be used to determine if suburbs with varying socioeconomic conditions have newer or older cars.

Codeset

Format (Example - 2010, 2012): “YYYY” is composed of 4 digits represented by Y.

Scope 1991-2023

Guidelines to Use

Must ensure that the entered year is valid.

Source

The manufactured year of the cars was derived using the AUCN API.

Age

Definition Age is an integer that denotes the age of the car.

Justification The year of a car can be used to determine if suburbs with varying socioeconomic conditions have newer or older cars.

Codeset **Format** Positive Integer (Examples: 0,1,2...) “99” is composed of 1 or 2 digits represented by “9”. The length may vary, but usually one or two digits, representing the car's age in full years.

Scope 0-31

Guidelines to Use Age must not contain any negative values.

Source Age was calculated by subtracting the current year (2023) from the manufactured year of the car.

Average Dealer Value

Definition The average dealer retail price of the car.

Justification This value is determined by various factors including brand, model, year, mileage, condition, market demand, and economic trends.

Codeset Average Dealer Value codeset includes numerical values representing currency units (AUD).

Format Fixed-Point Number (Examples: $45,460, $69,865).

Scope $2,230 - $504,025

Guidelines to Use To avoid ambiguity, it is essential to always indicate the currency alongside the value.

Source The average dealer retail price of the cars was derived using the AUCN API.

Score of Index of Relative Socio-economic Disadvantage

Definition The Score of the Index of Relative Socio-economic Disadvantage (IRSD) is a statistical tool used to evaluate an area's relative socioeconomic conditions.

Justification The data is derived from various factors including income, education level, employment status, and occupation of residents. A lower score indicates a greater disadvantage in access to resources.

Codeset	**Format** Numeric, with 6 decimal points for precision (Example: 1034.925695, 995.1950812) **Scope** 717.3364765 - 1097.262255
Guidelines to Use	Must be a within the defined scope
Source	Socio-economic Score was derived from the Australian Bureau of Statistics

We reverse-engineered the Australian Car Network (AUCN) website (AUCN, n.d.) to identify the source of their car data and discovered the API mentioned below,

https://api.aucn.net.au/aucar/valuation?rego=BJO48A&state=SA

A Python script was developed to collect car data using their API, focusing on manufacturing year and average dealership value for each car plate.

Appendix 02 - Data Quality Assessment

Timeliness

Timeliness refers to the extent to which the data accurately represent the reality of car number plates at the specific point in time when they were collected. Each field in the collected data set can be considered to discuss how each of them corresponds to timeliness.

Car Plates

Relation to Timeliness: A car plate on its own doesn't provide direct information about timeliness. However, if plates were recorded over a span of time, the date of collection becomes important. For instance, if car plates from affluent suburbs were primarily collected recently while those from less affluent suburbs were collected years ago, this can introduce a timeliness bias.

Selection Bias in Relation to Timeliness: There might be selection bias if plates from certain time periods or specific events were prioritised or ignored. For example, collecting data during a luxury car exhibition might skew the representation of a suburb's socio-economic status.

Postcode

Relation to Timeliness: The postcode can reveal the suburb's socio-economic status. Changes in the socioeconomic conditions of suburbs can happen over time. If data from a certain postcode is outdated, then it may not represent the current socio-economic status of that suburb.

Selection Bias in Relation to Timeliness: Bias may emerge if older data from specific postcodes is included while more recent data from other postcodes is not, or vice versa.

Year of Manufacturing of the Car

Relation to Timeliness: The year of manufacturing gives an idea about the age of the car, which can indirectly indicate the socio-economic status. Newer cars might suggest a higher socio-economic status. However, timeliness plays a role here. For example, If the dataset has older cars but those data points were collected many years ago, they might have been new at the time of collection.

Selection Bias in Relation to Timeliness: If the data collection focused more on newer cars and ignored older ones (or vice versa), there's a timeliness bias. This is because the emphasis on car age without considering the collection year might misrepresent the socio-economic conditions of the suburb at that time.

Age (from the car manufacturing date till the current year 2023)

Relation to Timeliness: This field essentially enhances the understanding of the previous field. By calculating the car's age up to 2023, it provides a direct measure to analyse the potential socio-economic status of the suburb in the present context.

Selection Bias in Relation to Timeliness: A bias would be evident if the dataset predominantly included cars of a certain age range. For example, if most cars are only 1-2 years old in the dataset, then the data might not be reflective of the broader socio-economic landscape of the suburb.

In summary, the concept of timeliness is crucial when assessing data quality. It's vital to understand when the data was collected and how relevant it remains in the current context. In the given scenario, while car plates and manufacturing dates provide indirect insights into timeliness, the age of the car (related to 2023) is a more direct measure. Any selection bias, such as favoring certain time periods or car ages, can compromise the data's ability to accurately reflect the socio-economic status of different suburbs.

Completeness

Completeness is achieved by capturing all types of car number plates in the suburbs, accounting for variations like different car number plate types (standard, personalised, specialty), and ensuring accuracy of the collected data. Each of these fields can be examined in the context of the concept of completeness during a data quality assessment.

Car Plates

Relation to Completeness: A dataset is deemed complete in this context if it has the car plate details for every vehicle surveyed. Missing car plate entries might make it difficult to cross-reference or validate data or identify unique vehicles.

Selection Bias in Relation to Completeness: If only specific types of car plates (for example, vanity plates or plates from a certain series) were recorded, and others were ignored, this would be a sign of selection bias affecting data completeness.

Postcode

Relation to Completeness: Having the postcode for each car is crucial. Missing postcodes mean it would be difficult to accurately tie a vehicle to a suburb, rendering its socio-economic inference moot. A complete dataset would have a valid postcode for every car plate recorded.

Selection Bias in Relation to Completeness: If data collection intentionally or unintentionally left out certain postcodes or focused excessively on others, it would introduce a bias. For instance, if data collectors skipped areas deemed as 'low-income', the data set would not be truly representative of the broader socio-economic landscape.

Year of Manufacturing of the Car

Relation to Completeness: The year of manufacturing is essential for inferring the age of the car, which is a proxy for socio-economic status. If this is missing for many entries, the dataset's ability to gauge socioeconomic status accurately would be compromised.

Selection Bias in Relation to Completeness: If cars from specific manufacturing years were more often recorded than others, or if cars from a certain range of years were ignored, this would introduce a completeness bias. For instance, ignoring older cars might give an inflated view of the socio-economic status.

Age (from the car manufacturing date till the current year 2023)

Relation to Completeness: This field, derived from the manufacturing year, provides a direct way to gauge the car's age. If the dataset does not calculate or provide this for each car, there is an issue of incompleteness, making the socio-economic inference process less efficient.

Selection Bias in Relation to Completeness: The bias here would largely arise from biases in the 'Year of Manufacturing' field. If certain age groups of cars were underrepresented or overrepresented due to biases in recording the manufacturing year, it would impact the completeness and accuracy of the dataset.

In summary, completeness in data quality assessment ensures that the dataset has all the necessary information for each entry and is representative of the actual scenario. In the given context, any missing data or over/underrepresentation in the fields can compromise the ability of the dataset to reflect the socio-economic status of different suburbs accurately. The potential for selection bias in relation to completeness exists in the process of data collection and entry, and it's crucial to be aware of these biases to ensure accurate interpretations and conclusions.

Uniqueness

Uniqueness in the context of the car number plate dataset signifies that no instance of a specific car number plate is recorded more than once within the dataset. The relationship of each field to the concept of uniqueness during a data quality assessment can be explored:

Car Plates

Relation to Uniqueness: The car plate should be a unique identifier for each vehicle. No two cars should have the same plate number. If there are duplicate plate entries, it could indicate data entry errors or other issues affecting data integrity.

Selection Bias in Relation to Uniqueness: If the data collection method favored repeated collection from certain areas or from the same vehicles, it might lead to non-unique car plate entries. For example, if cars in a high-end parking lot were surveyed multiple times, it could over-represent those plates in the data.

Postcode

Relation to Uniqueness: While postcodes won't be unique since multiple cars can come from the same suburb, a car plate should correspond to only one postcode. If the same car plate appears with different postcodes, it suggests inconsistencies in the data.

Selection Bias in Relation to Uniqueness: There may not be a direct uniqueness bias tied to postcodes. However, biases in other fields (like car plates) might indirectly lead to skewed representations of postcodes.

Year of Manufacturing of the Car

Relation to Uniqueness: This field will not be unique as many cars can be manufactured in the same year. But combined with the car plate, it can be used to detect inconsistencies. For instance, if the same car plate is associated with different manufacturing years in separate entries, it suggests data errors.

Selection Bias in Relation to Uniqueness: No direct uniqueness bias is anticipated for the manufacturing year. However, if there was a tendency to record certain manufacturing years repeatedly (perhaps due to a misunderstanding or misinterpretation), it could introduce redundancy in the dataset.

Age (from the car manufacturing date till the current year 2023)

Relation to Uniqueness: Similar to the year of manufacturing, this field won't be unique on its own since many cars will share the same age. Yet, in combination with the car plate, inconsistencies can be spotted. A single car plate should have one consistent age across entries.

Selection Bias in Relation to Uniqueness: Like the manufacturing year, there isn't a direct uniqueness bias for this field. However, indirect biases might arise from errors or repeated tendencies in other fields.

In summary, when it comes to the uniqueness criteria in data quality assessment, the primary focus should be on the car plates as they are expected to be distinct identifiers. The other fields, when cross-referenced with the car plates, help in identifying inconsistencies or duplications in the dataset. Any lack of uniqueness, especially with car plates, can compromise the accuracy and integrity of the data. Selection biases can indirectly lead to issues of uniqueness, especially if data is collected repeatedly from the same sources or if errors in data recording and entry are not rectified.

Validity

Validity, within the context of the car number plate dataset, implies the degree to which the data adheres to predetermined rules, standards, or specifications about the structure, category, and permissible values linked to car number plates. Each field concerning the concept of validity during a data quality assessment can be assessed

Car Plates

Relation to Validity: Validity for car plates means that each plate should adhere to the format and rules established by the issuing authority in the region. For example, in many places, car plates have a specific combination of letters and numbers. If a plate does not match the recognised format, it may be deemed invalid.

Selection Bias in Relation to Validity: If data collectors preferentially recorded plates from specific regions (with distinct plate formats) and neglected others, this can introduce a validity bias. Similarly, if fake or misrepresented plates were recorded, it would compromise validity.

Postcode

Relation to Validity: A valid postcode should be in line with the standardised postal coding system of the country. Any entry that doesn't adhere to the correct postcode format or lists non-existent postcodes would be invalid.

Selection Bias in Relation to Validity: If data collectors intentionally or accidentally focused on easily recognisable or familiar postcodes and avoided unfamiliar ones, it could introduce a bias. Also, recording postcodes without proper verification can lead to invalid data entries.

Year of Manufacturing of the Car

Relation to Validity: The year of manufacturing should be a realistic value. For instance, if the data set was collected in 2023, any car listed with a manufacturing year beyond 2023 would be invalid.

Selection Bias in Relation to Validity: If there was a preference for recording cars of a certain age (perhaps newer cars to reflect a higher socio-economic status), it might lead to overlooking older vehicles, introducing a validity bias in representing the socio-economic status accurately.

Age (from the car manufacturing date till the current year 2023)

Relation to Validity: The age should be a derived value from the year of manufacturing and should accurately reflect the car's age up to 2023. If there's a discrepancy between the manufacturing year and the provided age, the data entry might be invalid.

Selection Bias in Relation to Validity: Bias could arise if data collectors, intentionally or unintentionally, recorded or calculated the age inaccurately to skew the perception of the socio-economic status of a suburb. For instance, under-reporting the age to present cars as newer.

In summary, validity in data quality assessment ensures that the dataset entries are accurate, reliable, and adhere to established standards or formats. In this context, the validity of car plates and postcodes is about ensuring they conform to recognised formats, while the validity of the year of manufacturing and age is about ensuring realistic and consistent values. Selection biases can compromise validity by introducing preferences, errors, or misrepresentations during data collection, which in turn can impact the project's ability to accurately reflect the socio-economic status of different suburbs.

Accuracy

Data accuracy for the car number plate dataset refers to the extent to which the recorded number plates correctly represent the vehicles in the surveyed suburbs. It assesses the precision of the dataset in comparison to the actual number plates of the vehicles captured during the survey. Here's a breakdown of how each field relates to accuracy:

Car Plates

Relation to Accuracy: The car plate entries must precisely represent the actual number plates of the vehicles surveyed. An accurate dataset would have no typos, or other errors in the car plate data.

Selection Bias in Relation to Accuracy: If data collectors mistakenly recorded plates or copied them from unreliable sources, it would affect accuracy. Further, if they gravitated towards more legible plates and avoided unclear ones, it could introduce a bias in the representation of car plates, affecting the overall accuracy.

Postcode

Relation to Accuracy: The postcode must correctly correspond to the suburb from which the car originates. If there's a mismatch between the postcode and suburb, or if the postcode is entered incorrectly, it would impede accuracy.

Selection Bias in Relation to Accuracy: Collectors might be more familiar with certain postcodes and, hence, might record them with higher accuracy than unfamiliar ones. If there's a preference for specific suburbs or avoidance of others, this could lead to inaccuracies in how suburbs are represented.

Year of Manufacturing of the Car

Relation to Accuracy: The year of manufacturing should precisely match the actual manufacturing year of the car. Inaccuracies can arise if the data is guessed, estimated, or taken from unreliable sources.

Selection Bias in Relation to Accuracy: A bias might be introduced if data collectors, either due to ease or preference, predominantly record cars from certain years or avoid others. For instance, focusing on newer cars might give a skewed view of the socio-economic status.

Age (from the car manufacturing date till the current year 2023)

Relation to Accuracy: The age should be a straightforward derivation from the year of manufacturing, and it should be accurate. If the age doesn't align with the difference between the manufacturing year and 2023, there's an accuracy issue.

Selection Bias in Relation to Accuracy: If collectors consistently miscalculate age or base age on visual estimations rather than factual data, it could introduce inaccuracies. A preference to represent cars as newer or older (to reflect a certain socio-economic status) can also introduce bias.

In summary, accuracy in data quality assessment ensures that the data is a true representation of reality. Each field's accuracy is paramount to derive meaningful insights about the socio-economic status of different suburbs based on the cars' details. Selection biases, whether arising from preferences, ease of data collection, or misconceptions, can compromise this accuracy, leading to a distorted view of the socioeconomic landscape. **Consistency**

Consistency in a data quality assessment refers to the uniformity and reliability of data across the dataset. Data should be captured, represented, and processed in a consistent manner. Here's an examination of each field in relation to consistency: Car Plates

Relation to Consistency: Car plate entries should be consistently formatted throughout the dataset. Whether it's the use of hyphens, spaces, or the order of letters and numbers, the recording method should be uniform.

Selection Bias in Relation to Consistency: If data collectors have varying methods of recording car plates or if there are multiple sources of data with different recording standards, this could introduce inconsistency. For example, one collector might use spaces, while another uses hyphens.

Postcode of the Suburb

Relation to Consistency: The format and representation of postcodes should be consistent across entries. If a postcode system uses a fixed number of digits or characters, every entry should adhere to this.

Selection Bias in Relation to Consistency: Bias can arise if data collectors use different reference sources for postcodes or if they rely on memory for some and documentation for others. Inconsistencies might emerge if different collectors have different understandings of how to record postcodes.

Year of Manufacturing of the Car

Relation to Consistency: The year should always be represented in the same format, typically as a fourdigit number (e.g., 2020). If some entries use two digits (e.g., '20) while others use four, it leads to inconsistency.

Selection Bias in Relation to Consistency: If collectors, based on their discretion, decide to record the year differently (e.g., noting down "recent" instead of the exact year for newer cars), it could introduce inconsistency in the dataset.

Age (from the car manufacturing date till the current year 2023):

Relation to Consistency: The age should be consistently calculated as the difference between the current year (2023) and the manufacturing year. If there are different methods used, such as approximations or rounding off, it leads to inconsistency.

Selection Bias in Relation to Consistency: If some collectors estimate age visually (e.g., "looks 5 years old") while others calculate it based on the manufacturing year, it can lead to inconsistent data. Bias arises if there's no standardized method for determining age.

In summary, consistency ensures that data is recorded and represented uniformly across the dataset. Inconsistencies can lead to confusion, misinterpretation, and errors in analysis. In relation to the socioeconomic status project, inconsistent data can distort the understanding of the relationship between car details and socio-economic status. Selection biases that introduce inconsistency typically arise from nonstandardised data collection methods, varying understandings among collectors, or reliance on multiple and differing data sources.

Appendix 03 - Meeting Agenda and Quality Register

Task Assignment Table 04/11/2023

Person	Tasks Assigned	Quality Register
Bella's tasks	Material for analysis (Database explain), Analytic approach, Interpretations	☑
Alex's tasks	Background and Motivation (intro and justification), Material for analysis (Database explain), Analytic approach	☑
Maheshi's tasks	Material for analysis (Data quality), Recommendations	☑
Saif's tasks	Material for analysis (Data quality), Interpretations	☑
Sachini's tasks	Material for analysis (Data dictionary), Recommendations	☑
Moeez's tasks	Material for analysis (Data dictionary), Analytic approach, Recommendations	☑
Everyone's tasks	Findings	☑

Task Assignment Table 27/10/2023

Draft of report : 3th Nov Finilize on 4 Nov

Person	Tasks Assigned	Quality Register
Bella's tasks	Material for analysis (Database explain), Analytic approach, Interpretations	☑
Alex's tasks	Background and Motivation (intro and justification), Material for analysis (Database explain), Analytic approach	☑
Maheshi's tasks	Material for analysis (Data quality), Recommendations	☑

Saif's tasks	Material for analysis (Data quality), Interpretations	☑
Sachini's tasks	Material for analysis (Data dictionary), Recommendations	☑
Moeez's tasks	Material for analysis (Data dictionary), Analytic approach, Recommendations	☑
Everyone's tasks	Findings	Saif 30th Oct

Person	Tasks Assigned	Quality Register
Bella's tasks	Disadvantage points	☑
Alex's tasks	Conclusion	☑
Maheshi's tasks	Advantage	☑
Saif's tasks	Disadvantage example	☑
Sachini's tasks	Advantage points	☑
Moeez's tasks	Introduction	☑

Next meeting MONDAY 10:00 pm, Zoom meeting.

3rd meeting will be on Wednesday, please come early, 11:00 am (in person practice)

Week11 Meeting Preparatior Due at 12pm 18 Oct Goup meeting: 18 Oct in class	Alex	Bella	Maheshi	Moeez	Sachini	Saif
Have you read all the assignment requirements? YES/NO	Yes	Yes		Yes		

Which Topic do you prefer? LeaveNUMBER e.g Topic 1 & Why? 1. Artificial meat - possible impact on Australian agribusiness (farmers) 2. Impact of AI on middle-sized business in Australia (or in the country of your choice) 3. Expansion of trade in national currencies (euro, yuan...) - impact on US dollar (and AUKUS economy) 4. What can be learned from COVID - impact on economy and society (compare countries with tough restricitons and those wiothout) 5. What is "stakeholder capitalism" and what it means for small/medium businesses and consumers 6. What is CBDC (Central bank Digital Currency) and how it will influecne small/medium business and citizens 7. Perspectives of electromobility - pros and cons; economic (and physical) challenges		Topic2 For our presentation on the impact of AI on middle-sized businesses in Australia, we can split it into two parts: opportunities and threats. We'll have 1 person introduce, 2 people talk about the opportunities, 2 people discuss the threats, 1 person wrap up with a conclusion.				2 The reason is because most middle-sized businesses are planning to downsize their employees because of AI technologies. I noticed this abrupt change in graph designing studios where people are considering using AI art generators which is taking away jobs from graphic designers. Also, content writing industry is vastly challenged because of AI text creation.
Background and Motivation (intro and justification)		Alex		Integrated All parts what've done on 18 Oct group meeting		
Material for analysis (database explain, data source and data quality, data dictionary)		Database explain:Bella + Alex, Data quality: Maheshi+Saif, Data dictionary: Sachini + Moeez		4Oct-11Oct Integrated All parts what've done on 18 Oct group meeting		
Analytic approach		Bella+ Alex+ Moeez		4Oct-18Oct Integrated All parts what've done on 18 Oct group meeting		
Findings		Everyone needs to prepare for this part		11Oct-18Oct Integrated All parts what've done on 18 Oct group meeting		
Interpretations		Bella + Saif		11Oct-18Oct Integrated All parts what've done on 18 Oct group meeting		

Recommendations	Sachini+ Maheshi+ Moeez	18Oct-25Oct

Week9 Meeting Preparation Due at 12pm 4 Oct	Alex	Bella	Maheshi	Moeez	Sachini	Saif
Have you read all the assignment requirements? YES/NO	Yes	Yes	Yes	Yes	Yes	Yes
What's the outline of the final report?	As recommended by the tutor	Background and Motivation (intro and justification) Material for analysis (database explain, data source and data quality, data dictionary) Analytic approach Findings Interpretations Recommendations	Background and motivation Material for analysis Analytic approach Findings Interpretations Recommendations	Background and motivation Material for analysis Analytic approach Findings Interpretations Recommendations	Background and motivation Material for analysis Analytic approach Findings Interpretations Recommendations	Background context and research approach Data collection for analysis, data quality assessment, data dictionary Data analysis method Research findings Evaluation Conclusion
Which part will u work on?	I'm most comfortable with data manipulation, data interpretation, recommendation, report design	Findings, or any part is good for me.	I suggest to work in the way we divided in the plan	Analytical plan , findings and recommendations.	I will work on evaluating the results from a business perspective and analyzing data.	Evaluation and reporting
Any other proposals?		I propose that Sachini works very well with Tableau; Sachini might work on Data Visualization. Conducting a data quality test is quite challenging, especially due to the bias in data collection. For instance, we found 962 duplicate plates, as 5 groups have the same reach area as Mawson Lakes. Additionally, cars moving from one suburb to another resulted in the	I propose everyone to work on the data analysis part so that we can select the best one for the final submission	I propose everyone to work on the data analysis part so that we can select the best one for the final submission	I propose that everyone evaluates the data and find the best way to interpret it.	I propose we fix data quality assessment first as per the recommendation given by Jane. I also want everyone to find out about different approaches we can take for data interpretation and analysis. Use descriptive statistics, histograms, box plots, and scatter plots to explore the distribution of car age and estimated car value in different suburbs. This will give you a preliminary understanding of the data. Calculate correlation coefficients (e.g., Pearson correlation) to quantify the relationship between car age and estimated car value. Determine if there's a significant correlation between these two variables. Perform regression analysis, such as linear regression, to model the relationship between car age and estimated car value. You can create separate models for each suburb to see if the relationship varies by

		same collected plate from different suburbs. We need to discuss all these biases.				location. Conduct hypothesis tests (e.g., t-tests or ANOVA) to determine if there are statistically significant differences among suburbs. Use storytelling techniques to convey insights and recommendations
Background and Motivation (intro and justification)		Alex		4Oct-11Oct		
Material for analysis (database explain, data source and data quality, data dictionary)		Database explain:Bella + Alex, Data quality: Maheshi+Saif, Data dictionary: Sachini + Moeez		4Oct-11Oct		
Analytic approach		Bella+ Alex+ Moeez		4Oct-18Oct		
Findings		Everyone needs to prepare for this part		11Oct-18Oct		
Interpretations		Bella + Saif		11Oct-18Oct		
Recommendations		Sachini+ Maheshi+ Moeez		18Oct-25Oct		

Week 9 Group Meeting Minutes 4 Oct 23	
1. Solid outline of final report	
2. Writing Allocation	This week: Draft of P1,P2 and P3 Next week: Modify and feedback the P1,P2 and P3. Two week later: Findings
3. Next meeting	11 Oct in class Discuss Data quality and other draft
4. Summary (5mins)	

Week 9 Group Meeting Agenda 4 Oct 23	
1. Solid outline of final report	
2. Writing Allocation	
3. Next meeting	Week 10 in class Discuss Data quality and other draft
4. Summary (5mins)	

Week 8 Group Meeting Agenda 17 Sep 23	
1.Warm up 15 mins	1. What’s Ur strength in group work? 2. What’s Ur weakness in group work? 3. What’s Ur hobbies?
2. Introduction of the Final Group assignment 5mins	Plan: 1. Finish Data Analysis before Week 9 2. Week 9 Start to write final report 3. Week 10 Finish Final report 4. Week 11 Submit
3. Discuss about the plan	If you agree or not? Why?
4. Summary (5mins)	