COMP90024 Assignment 2
Big Data Analytics on the Cloud

Group 53

- Parsa Babadi Noroozi (1271605)
- Niket Singla (1288512)
- Jason Phan (1180106)
- Patipan Rochanapon (1117537)
- Liam Brennan (1269948)

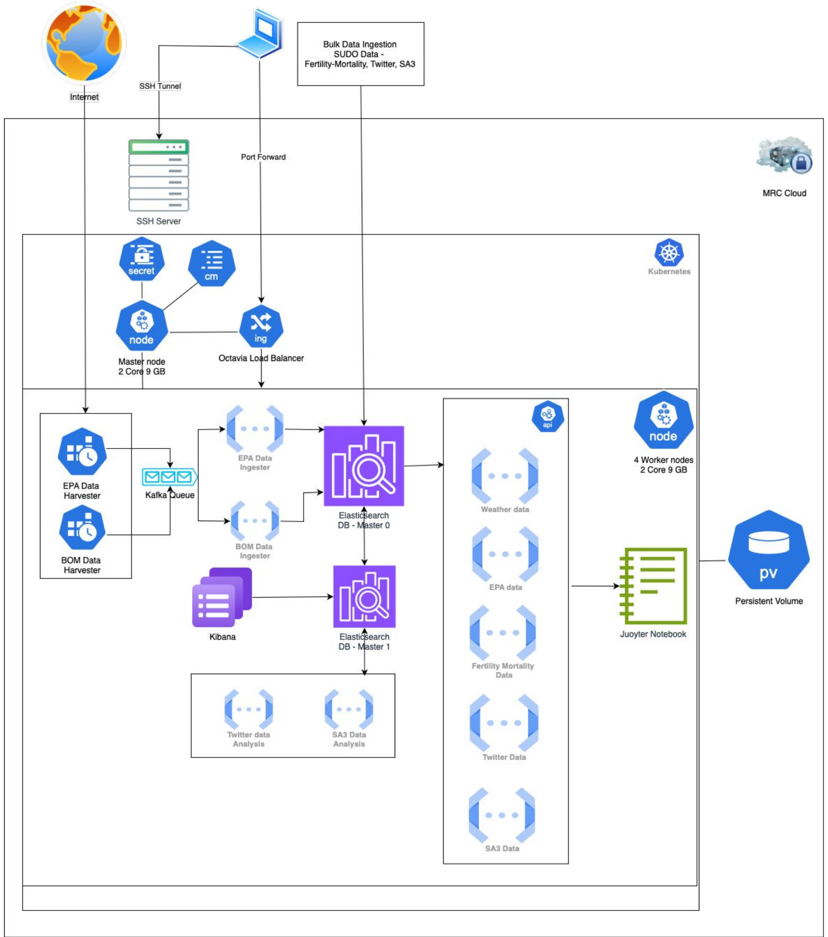
Team 53 Introduction

Niket Singla 1288512	Liam Brennan 1269948	Parsa Babadi Noroozi 1271605	Patipan Rochanapon 1117537	Jason Phan 1180106
• Cluster setup • Jupyter hub setup on K8 cluster • Kafka setup • EPA & BOM data ingestion • Elasticsearch • Data analysis API • Jupyter frontend • Unit test for all ingestion function and API endpoints	• Crash and health risk processing backend • Joined crash and SA2 analysis API • Crash and health risk frontend visualisation • Unit and end-to-end API endpoint testing	• Crash and health risk data collection • Crash and health risk elasticsearch ingestion • Crash and health risk front end visualisation • Crash and health risk analysis	• Twitter, SA3, SUDO (SA3) Ingestion • Twitter and SA3 coordinates mapping • Join Twitter with SUDO (SA3) • Twitter and SA3 Data Analysis	• Twitter API routes • SA3 data API routes • Twitter and SA3 Data Analysis • Twitter sentiment and age/income/education data visualisations • Elasticsearch queries • End to End testing for API endpoints

System Architecture

Following are the major components of our project deployed on MRC

- Cloud infrastructure, including necessary RAM, storage and processing capacity
- Kubernetes (K8s)
- Kafka
- Fission
- Elasticsearch & Kibana
- Jupyter notebook



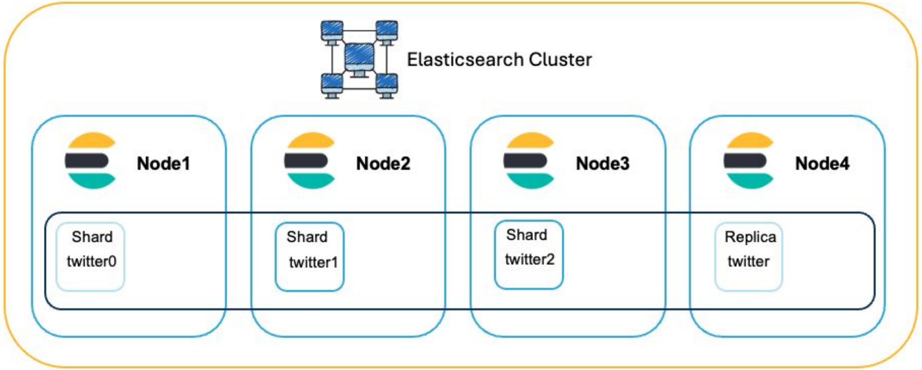
Elasticsearch

Indexing: Efficient storage and retrieval of data.

Sharding & Replication: Distribution and replication of data across nodes for scalability and fault tolerance.

Ingest Pipelines: Preprocessing data before indexing using custom pipelines for parsing, enriching, and modifying documents.

Querying and Analysis: Retrieving and analyzing data using the Elasticsearch query DSL.



Elasticsearch Cluster

Node1

Shard twitter0

Node2

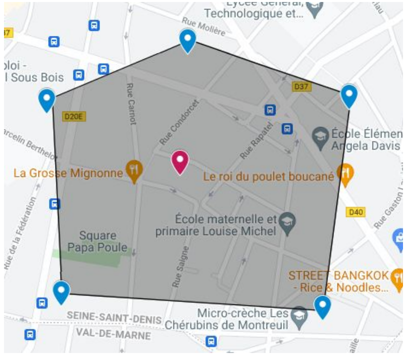
Shard twitter1

Node3

Shard twitter2

Node4

Replica twitter



Data Collection

Data Collection via API

- Environment Protection Authority Victoria (EPA)
- Bureau of Meteorology (BOM)

SUDO Dataset

- SA3 Population, Highest Education, and Average Age & Income
- SA2 Health Risk Factors
- Combined SLA11 Premature Mortality & Fertility

Other External Datasets

- Crash Dataset
- SA2 & SA3 GeoJson Dataset
- Twitter Dataset

Backend

Twitter and SUDO (SA3)

- Mapping Twitter coordinates with SA3 polygons.
- Utilize Elasticsearch query with "geo_distance" filter to pinpoint nearest SA3 polygons for each Twitter coordinate.
- Update Twitter index with "sa3_code_2021" to exclude non-Australian coordinates.
- Join Twitter and SUDO index by sa3_code_2021

SA3

Join

Twitter

Road Crashes and SA2 Health Risks

- Road crash geolocations are joined using intersection queries against SA2 district geometries
- Corresponding SA2 health risk data is incorporated as part of the joining process
- The final joined dataset combines crashes, health risks, and SA2 information together
- This dataset can be analysed using grouping and metric aggregation queries to extract insights

Crashes

Data Analysis - Research Questions

Weather & Fertility/Mortality Rates

- Is there a relationship between air quality and temperature?
- Is there a relationship air quality, weather and mortality and fertility?

Crash & Alcohol Consumption

- How are car crashes distributed geographically within Victoria?
- How do risk factors such as alcohol consumption affect crash severity?
- Explore the densities of alcohol consumption and car crashes

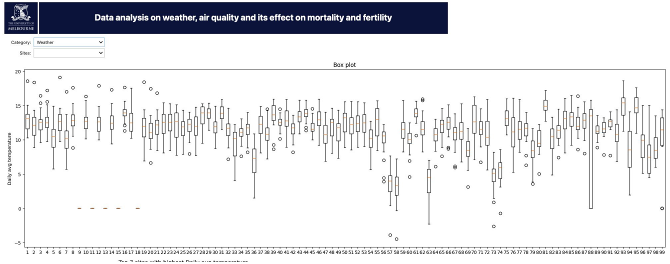
Twitter

- Twitter Use vs. Population Counts
- Is there a relationship between happiness and income levels?
- Is there a relationship between happiness and education?
- Is there a relationship between happiness and age?

Questions

Thank you

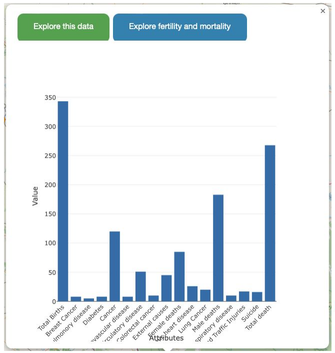
Weather, Air Quality and Fertility-Mortality Data Analysis

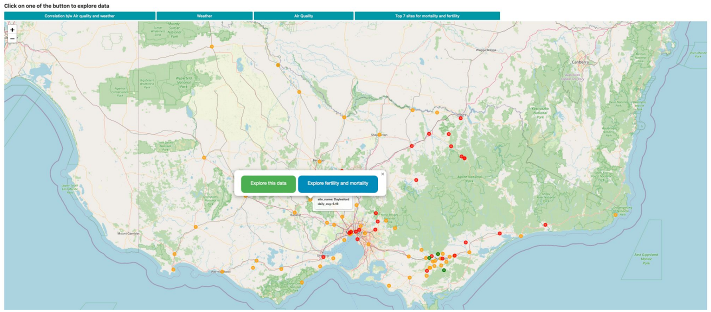


Data analysis on weather, air quality and its effect on mortality and fertility

Category: Weather

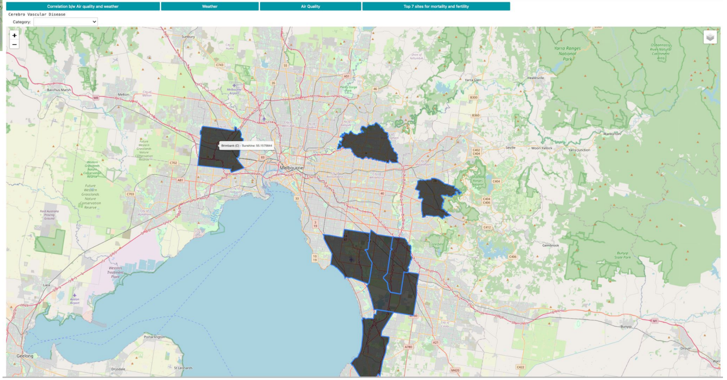
Sites:





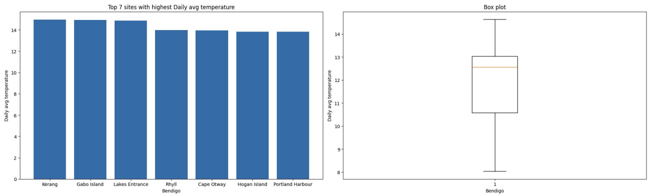
Click on one of the button to explore data

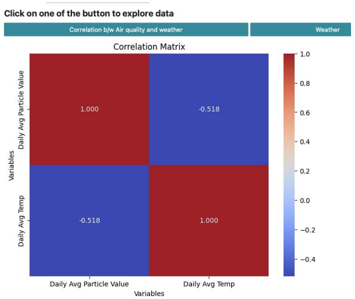
Correlation b/w Air quality and weather
Weather
Air Quality
Top 7 sites for mortality and fertility
+
−
Explore this data
Explore fertility and mortality
Melbourne
Canberra



Category: Cerebro Vascular Disease

Correlation b/w Air quality and weather
Weather
Air Quality
Top 7 sites for mortality and fertility
+
−
Melbourne
Geelong



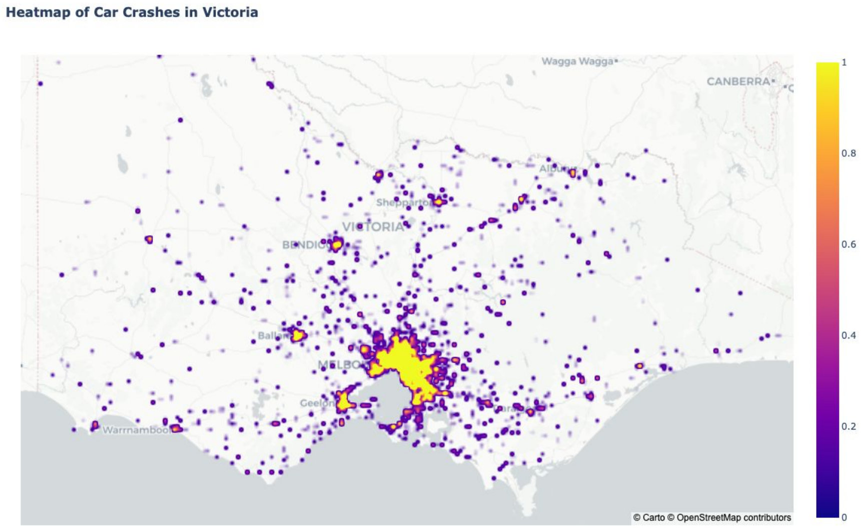


Click on one of the button to explore data

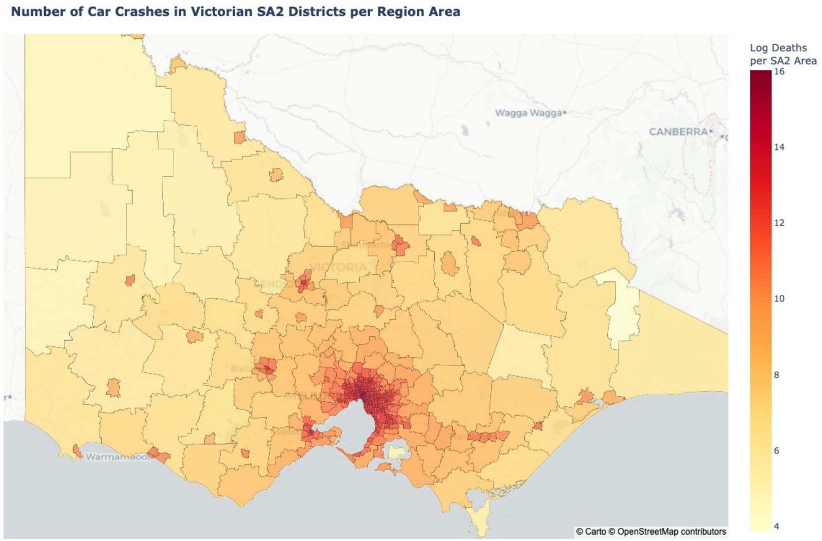
Road Crashes and SA2 Health Risks

Exploring Crash Incidence

Swap right map out with new one after demo practice


Heatmap of Car Crashes in Victoria

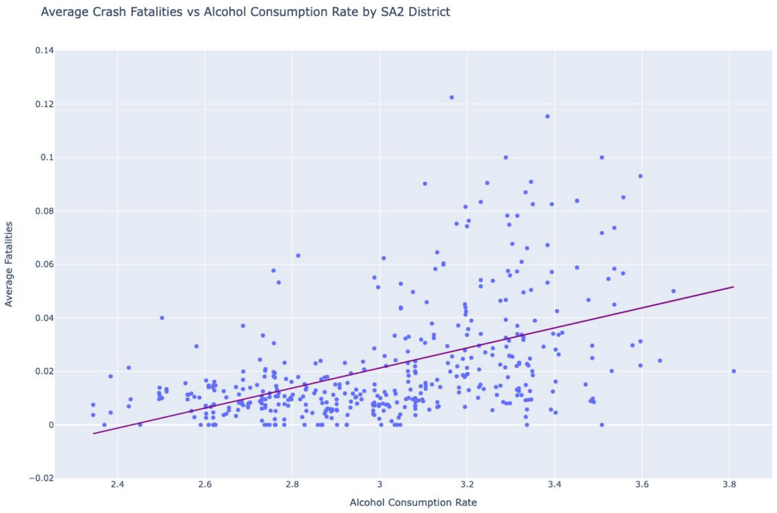
Distribution of car crash locations


Number of Car Crashes in Victorian SA2 Districts per Region Area

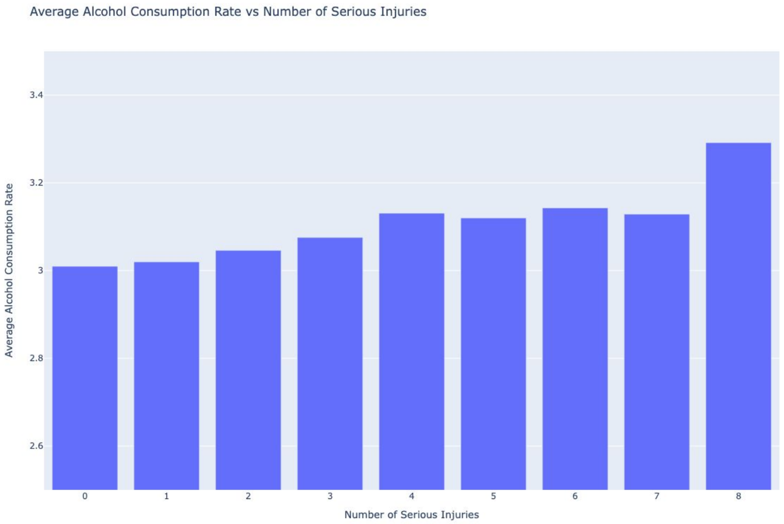
SA2 region crash incidence comparison

Road Crashes and SA2 Health Risks

Exploring Health Risk Influences


Average Crash Fatalities vs Alcohol Consumption Rate by SA2 District

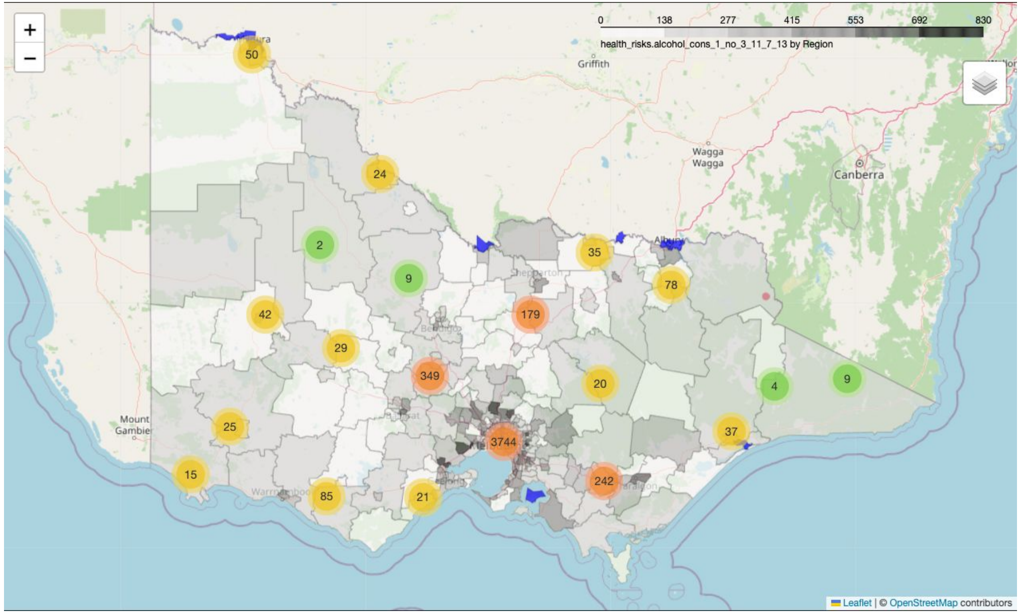
Crash severity (approximated by fatalities) vs alcohol consumption rate


Average Alcohol Consumption Rate vs Number of Serious Injuries

Alcohol consumption rate vs crash severity (approximated by number of serious injuries)

Road Crashes and SA2 Health Risks

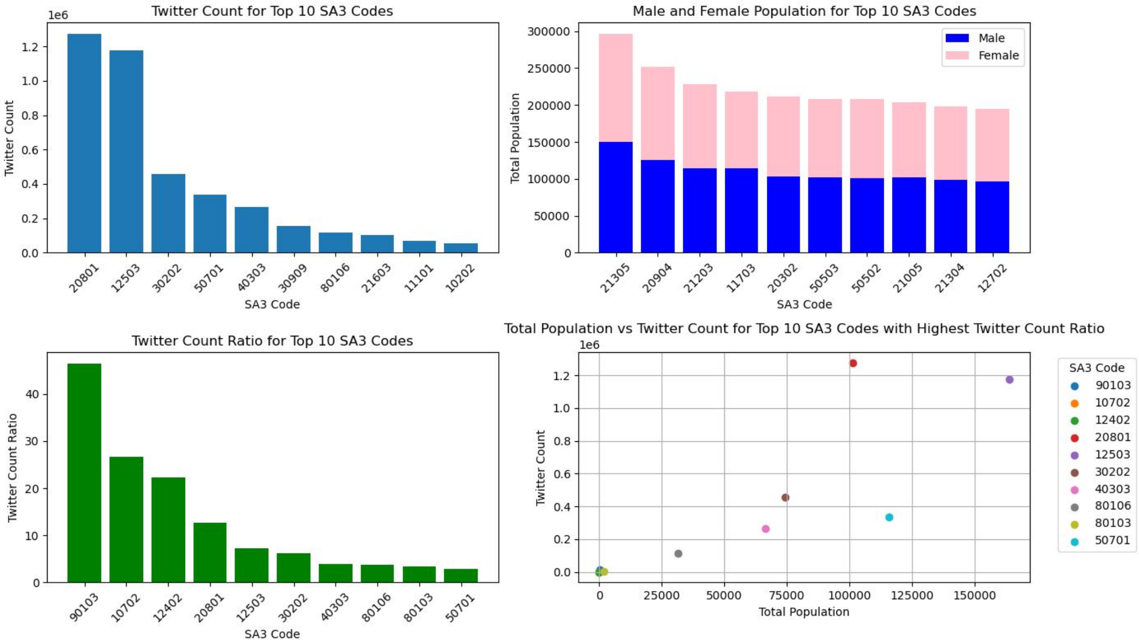
Exploring Health Risk Influences



SA2 alcohol consumption counts vs car crash distribution

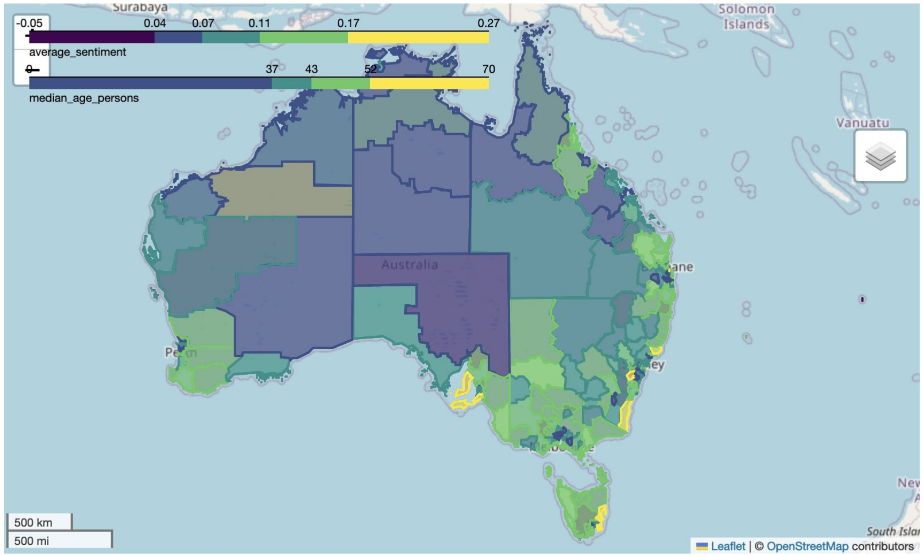
Twitter Sentiment Analysis

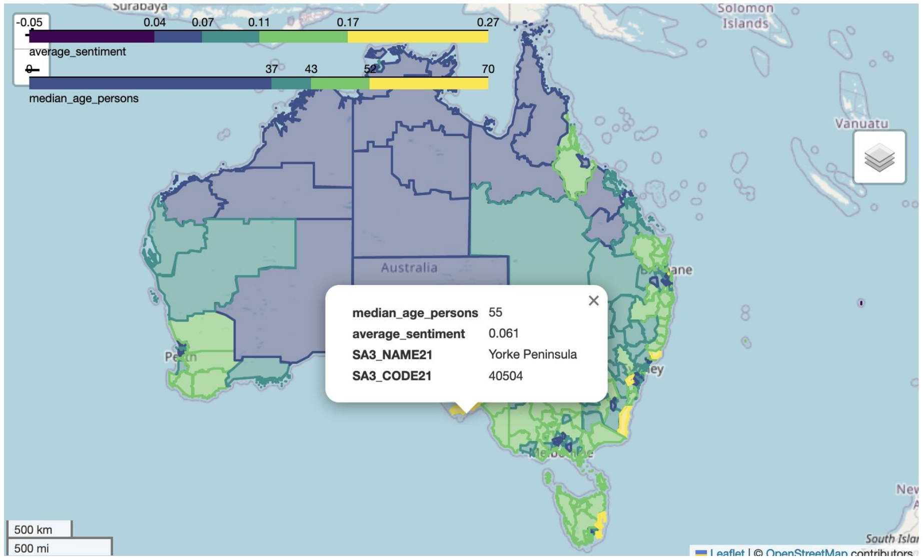
Twitter Count VS Population



Twitter Sentiment Analysis

Is there a relationship between happiness and age?

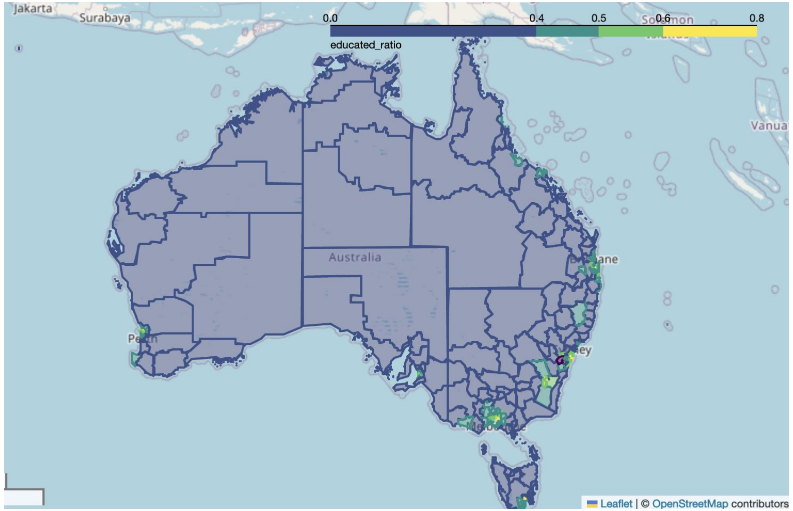


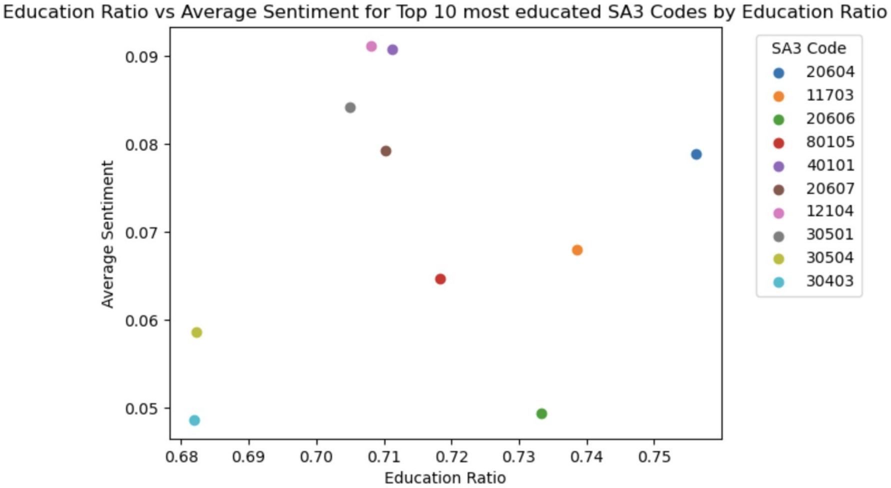


median_age_persons	55
average_sentiment	0.061
SA3_NAME21	Yorke Peninsula
SA3_CODE21	40504

Twitter Sentiment Analysis

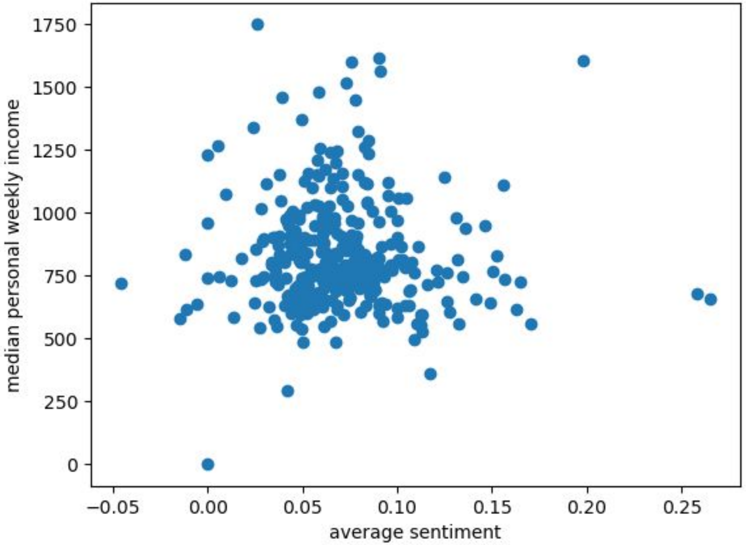
Is there a relationship between happiness and education?





Twitter Sentiment Analysis

Is there a relationship between happiness and income levels?


median personal weekly income
average sentiment