**Google Collab**

Numpy

Pandas

Apache beam pipeline - streaming

* Stream data
* Explore data
* Clean data
* Prepare data
* Modeling
* Evaluate
* Visualize results

Datasets:

<https://www.statista.com/statistics/971696/annual-loans-of-physical-library-materials-singapore/>

<https://www.statista.com/statistics/1228203/singapore-digital-interaction-rankings-public-sector-organizations/>

<https://www.statista.com/statistics/877309/singapore-coffee-consumption-volume/>

What factors affect an athletes performance the most - equipment / training / external factors?

<https://www.kaggle.com/datasets/coni57/f1-2020-race-data>

* Predict race timings

<https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires>

can predict

-fire size

-high risk areas

-cause of fires

<https://www.kaggle.com/datasets/ulrikthygepedersen/crossfit-athletes>

Predict athlete performance

<https://www.kaggle.com/datasets/aiaiaidavid/ironman-703-race-data-between-2004-and-2020>

* Predict overall shortest timing

Olympic Data

<https://www.kaggle.com/code/chadalee/olympics-data-cleaning-exploration-prediction/input?select=athlete_events.csv>

* Predict athlete playing career

<https://www.kaggle.com/datasets/maryalebron/life-expectancy-data>

-predict life expectancy/mortality rates/health expenditure

<https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_geolocation_dataset.csv>

Tourism:

<https://www.kaggle.com/datasets/matarrgaye/uk-consumer-trends-current-price>

<https://www.kaggle.com/datasets/dasmehdixtr/tourist-number-data-in-turkey>

Crime

Social media influence

Price of traveling in country

Seasonal events

<https://www.kaggle.com/datasets/dipeshkhemani/airbnb-cleaned-europe-dataset>

Airbnb:

What factors affect price of airbnb

* Room type
* Country of airbnb
* Location within country
* Demand = holiday season / events (e.g. festival / sport event)

<https://www.kaggle.com/datasets/thedevastator/airbnb-price-determinants-in-europe>

<https://data.opendatasoft.com/explore/dataset/airbnb-averages%40public/export/?disjunctive.room_type>

<https://github.com/nit611/airbnb_singapore/blob/master/listings.csv>

[Inside Airbnb: Get the Data](http://insideairbnb.com/get-the-data/)

[Airbnb Prices in European Cities | Kaggle](https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities) - 20k rows

– GOAT

<https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata> - 35 million rows

<https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data> - alot of rows

<https://www.kaggle.com/datasets/davidcariboo/player-scores?select=appearances.csv>

<https://www.kaggle.com/datasets/efehandanisman/skytrax-airline-reviews> - reviews

<https://www.kaggle.com/datasets/rakshitrk/airline-reviews> - reviews

**Problem**

Airline customers are increasingly demanding a high-quality travel experience. Airlines need to be able to **predict the quality of their service** in order to improve customer satisfaction and loyalty.

Airlines Predict

* Price - what factors determine price **zowie**
* Delay - how does weather affect severity of delay **siddarth**
* Airport Review - is departing / arriving airport a factor in quality of flight **zx**
* Airlines reviews - overall which airline has the best performance **parik**

Data - 2022

- depart from / arrive at LAX

- reviews of respective airlines

- reviews of lax

- weather at us

<https://www.kaggle.com/datasets/sherrytp/airline-delay-analysis> -

<https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022>

```

<https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022?select=Combined_Flights_2018.csv>

<https://www.kaggle.com/datasets/timmofeyy/all-the-airport-across-the-world-dataset>

<https://www.kaggle.com/code/rajeevnair676/airline-satisfaction-xgboost-96-accuracy/input?select=train.csv> - satisfaction

<https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay>

<https://www.kaggle.com/datasets/saadsikander/tweet-for-airline> - reviews

<https://www.kaggle.com/code/divyansh22/airline-reviews-eda-and-preprocessing-pt-1/input> - reviews

<https://www.kaggle.com/datasets/thedevastator/sentiment-analysis-of-us-airline-twitter-data> - us airline reviews

Scrape lax reviews - google reviews

<https://www.kaggle.com/datasets/dilwong/flightprices> -5mil

<https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events> - weather dataset from 2016 -2022

<https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations> - weather 5 mil

Airport review website

<https://www.airlinequality.com/review-pages/a-z-airport-reviews/>

How can we use machine learning to **predict the optimal price for selling a residential property** in Singapore based on various factors such as location, size, type, amenities, and market conditions?

Some possible sub problems that are not related to the technical process are:

* How does the location of the residential property affect the price and the demand in Singapore?
* How does the size of the residential property influence the price and the comfort in Singapore?
* How does the type of the residential property determine the price and the preference in Singapore?
* How does the amenities of the residential property impact the price and the satisfaction in Singapore?
* How does the market conditions of the residential property vary by time, season, and economic factors in Singapore?

Topic:

Understand Project Scope & Expectation (For ZX own brainstorming)

1. Try something relate to Singapore
2. Something that helps a specific company, Find insights or help the government. It should include
3. Looking for a dataset that answers a problem statement(hypothesis? Example: Is Singapore a good place to live in? Then work on finding traffic, education, consumer goods, security and so on).
4. With the dataset it will be able to perform big data streaming.
5. Able to build models and make predictions using dataset.

Problem Statement: The availability of affordable housing in different regions of Singapore significantly impacts the overall quality of life and housing satisfaction.

To investigate this hypothesis, we would need a dataset that includes information on housing prices, infrastructure, amenities, and public services in various neighborhoods across Singapore. This dataset will allow us to perform big data streaming and build models to make predictions related to housing affordability and quality of life.

Housing Prices: This section of the dataset would include historical and real-time data on housing prices in different regions of Singapore. It would cover various property types such as apartments, condominiums, landed houses, and HDB (Housing Development Board) flats. The data would include factors like location, size, number of rooms, and pricing trends.

Infrastructure and Amenities: This part of the dataset would contain information on the infrastructure and amenities available in different neighborhoods. It would include data on transportation networks (MRT stations, bus stops), accessibility to schools, healthcare facilities, shopping malls, recreational parks, and other essential amenities. This data will help assess the overall convenience and livability of each region.

Public Services: This section would focus on public services such as security, healthcare, and education. It would include data on crime rates, police stations, hospitals, clinics, and schools in different regions. Analyzing this data will provide insights into the safety, healthcare access, and educational opportunities available in each area.

Demographic and Socioeconomic Factors: This part of the dataset would encompass demographic information such as population density, age distribution, income levels, and employment rates in different regions of Singapore. By analyzing this data, we can gain insights into the socioeconomic background and lifestyle preferences of residents, which may impact housing choices and satisfaction.

Hypothesis/Problem Statement:

"The availability of green spaces and proximity to nature positively impacts the mental well-being and overall happiness of residents in urban areas."

Retail and E-commerce: Big data is used by retailers and e-commerce companies to gain insights into customer behavior, preferences, and purchasing patterns. It enables personalized marketing campaigns, targeted recommendations, inventory optimization, dynamic pricing, and demand forecasting, leading to improved customer experiences and operational efficiency.

Transportation and Logistics: Big data analytics is applied in the transportation and logistics industry to optimize routes, manage fleets, predict maintenance needs, and improve supply chain operations. It involves analyzing data from GPS systems, sensors, telematics, weather forecasts, and historical logistics data to optimize logistics networks, reduce costs, and enhance overall efficiency.

1. Fraud Detection in Financial Transactions:

Hypothesis: The analysis of transaction patterns and anomaly detection techniques can lead to the development of an effective fraud detection system that identifies suspicious activities in financial transactions.

2. Sentiment Analysis in Social Media Data:

Hypothesis: By applying natural language processing techniques to social media data, it is possible to determine the sentiment and opinions of users towards a specific product, brand, or topic.

3. Predicting Equipment Failure in Industrial Settings:

Hypothesis: Through the analysis of sensor data and historical maintenance records, a predictive model can be developed to identify equipment failures in industrial settings, enabling proactive maintenance and reducing downtime.

4. Optimizing Energy Consumption in Buildings:

Hypothesis: The analysis of energy usage patterns and building sensor data can lead to the development of an optimization model that identifies energy-saving opportunities and provides recommendations for reducing energy consumption in buildings.

5. Traffic Flow Prediction for Smart City Planning:

Hypothesis: By analyzing historical traffic data, weather conditions, and using time-series forecasting techniques, a predictive model can be developed to forecast traffic flow patterns, contributing to smart city planning and the optimization of transportation infrastructure.