Data Science Final Project Group 6

**Presentation Notes**: *Introduction*

**Purpose**

The purpose of this project is to demonstrate the skills we developed over our six month data visualization bootcamp class.

**Presentation**:

Google Slide Link: [Google Presentation](https://docs.google.com/presentation/d/15mORnZ5z8yIWXXm1Pm6wAKRais7dHYS29vaFD6OZ9b4/edit?usp=sharing)

**Team Members**:

* Liliia Strukova
* Silvia Eiden
* Hina Mahboob
* Maureen Hubka

**Selected topic:**

Travel insurance claims.

**Reason:**

The travel insurance claims dataset was chosen to investigate business opportunities to sell travel insurance. We wanted to investigate the pros and cons of investing in this type of business and gain insight into the market's trends. We used Machine Learning to gauge the accuracy of the dataset to see if it can be used to develop a business model plan.

**Description of the Data:**

The Travel Insurance Dataset is from a third-party travel insurance servicing company that is based in Singapore. The dataset was downloaded from [Kaggle](https://www.kaggle.com/datasets/mhdzahier/travel-insurance) website.

* The following attributes are contained in the dataset:
* Target: Claim Status (Claim.Status)
* Name of agency (Agency)
* Type of travel insurance agencies (Agency.Type)
* Distribution channel of travel insurance agencies (Distribution.Channel)
* Name of the travel insurance products (Product.Name)
* Duration of travel (Duration)
* Destination of travel (Destination)
* Amount of sales of travel insurance policies (Net.Sales)
* Commission received for travel insurance agency (Commission)
* Gender of insured (Gender)
* Age of insured (Age)

**Questions we looked to answer:**

* Is there a relationship between insurance claims and travel destination?
* Is there a relationship between insurance claims and type of travel insurance agencies?
* Is there a relationship between insurance claims and duration of travel?
* Is there a relationship between insurance claims and age of the insured?
* Can these data points be used to accurately predict insurance claims?
* Does type of sales (i.e.: online, product) affect claims or the amount of insurance sold?
* Are there any other trends that could be used to determine the likelihood of insurance claims being filed?

**GitHub**

A GitHub repository was started for the team to contribute and collaborate based on the portion of the final project being worked on. Each team member was to have at least four commits per segment to keep track of the work that had to be done. Below is a list of six branches created for the Data Science Final Project Group 6.

* Branches:
  + Main
  + Dashboard
  + Database
  + Machine Learning Model
  + Presentation
  + Requirements

**Database**

The Travel Insurance Claims Dataset csv file was downloaded from Kaggle. It was visually analyzed to determine what fields would be most useful. The dataset was cleaned with Pandas in Jupyter Notebook as followed:

* Gender column was dropped due the large number of null values.
* Claim column transformed to 1 for “yes” (insurance claimed) and 0 for “no”.
* There were a few rows of data with negative and zero values in the Duration column which were dropped.
* Several rows with duration longer than 600 days were dropped - all huge outliers.
* As per the boxplot for Age column the biggest outlier was 118 years - those rows were dropped.
* Net Sales and Commission columns were removed to a new DataFrame so our working DataFrame wouldn't be overfitted with unnecessary information.

The cleaned dataset was then imported to PostgreSQL. For these purposes `sqlalchemy` module was used to create the database engine to connect Pandas with SQL. Connection to PostgreSQL was also added to the Jupyter Notebook with Machine Learning Model so the dataset could be imported for further analysis. [Quick DBD](https://www.quickdatabasediagrams.com/) online tool was used to create ERD with relationships.