Rakamin Academy

Homework EDA

Final Project - Stage 1

by Kelompok 4, Byte Blazer:
- Muhamad Faiz Widagdo
- Robiatul Adawiyah
- Chianti Ridhwan
- Lulu Safira
- Retno Debbi Yulisya
- Imam Luthfi
- Melliza Nastasia Izazi

Data Cleansing

Data cleansing dilakukan untuk mendeteksi kesalahan pada data. Pada tahap ini kami melakukan pengecekan tipe data sesuai dengan kebutuhan, dan mengecek apakah ada data yang kosong dan duplicate.

A. Pengecekan Type Data

Untuk mengecek tipe data digunakan .info() , dari data tersebut kami mengubah type data pada kolom 'ChronicDiseas' dan 'TravelInsurance' dari int64 menjadi object (Yes/No).

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1987 entries, 0 to 1986
Data columns (total 10 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Unnamed: 0           1987 non-null   int64
 1   Age                  1987 non-null   int64
 2   Employment Type      1987 non-null   object
 3   GraduateOrNot        1987 non-null   object
 4   AnnualIncome         1987 non-null   int64
 5   FamilyMembers        1987 non-null   int64
 6   ChronicDiseases      1987 non-null   int64
 7   FrequentFlyer        1987 non-null   object
 8   EverTravelledAbroad  1987 non-null   object
 9   TravelInsurance      1987 non-null   int64
dtypes: int64(6), object(4)
memory usage: 155.4+ KB
```

```
df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1249 entries, 0 to 1985
Data columns (total 10 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Age                  1249 non-null   int64
 1   Employment Type      1249 non-null   object
 2   GraduateOrNot        1249 non-null   object
 3   AnnualIncome         1249 non-null   int64
 4   FamilyMembers        1249 non-null   int64
 5   ChronicDiseases      1249 non-null   object
 6   FrequentFlyer        1249 non-null   object
 7   EverTravelledAbroad  1249 non-null   object
 8   TravelInsurance      1249 non-null   object
 9   CatAnIncome          1249 non-null   category
dtypes: category(1), int64(3), object(6)
memory usage: 98.9+ KB
```

Data Cleansing

B. Pengecekan Data Kosong

```
[ ]  df.isna().sum()
```

```
Age                    0
Employment Type        0
GraduateOrNot          0
AnnualIncome           0
FamilyMembers          0
ChronicDiseases        0
FrequentFlyer          0
EverTravelledAbroad    0
TravelInsurance        0
dtype: int64
```

Selanjutnya untuk mengecek data yang kosong, menggunakan df.isna().sum() .Dari hasil tersebut ditemukan bahwa tidak ada data yang null.

EDA, Insights & Visualization

C. Descriptive Statistics - Tidak ada kolom yang memiliki nilai summary yang aneh

Kami mengelompokkan kolom menjadi Categorical dan Numeric

```
[ ] # pengelompokan berdasar jenis untuk melihat sari sisi numeric dan kategori
    cat = ['Employment Type','GraduateOrNot','FrequentFlyer','EverTravelledAbroad','TravelInsurance','ChronicDiseases']
    num = ['Age', 'AnnualIncome','FamilyMembers']
```

```
[ ] df.describe()
```

	Age	AnnualIncome	FamilyMembers
count	1249.000000	1.249000e+03	1249.000000
mean	29.755805	9.345476e+05	4.890312
std	2.921039	3.607293e+05	1.762313
min	25.000000	3.000000e+05	2.000000
25%	28.000000	6.000000e+05	4.000000
50%	29.000000	9.000000e+05	5.000000
75%	32.000000	1.200000e+06	6.000000
max	35.000000	1.800000e+06	9.000000

Pengamatan Numerical:

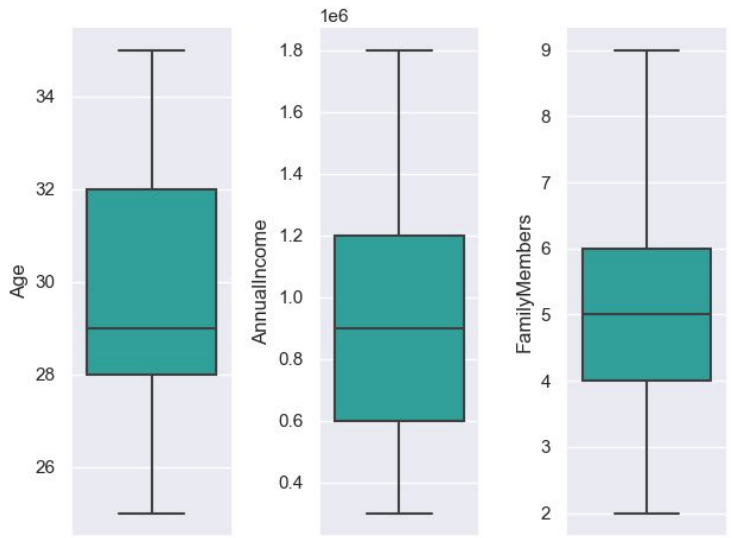
Ada perbedaan antara nilai rata-rata dengan median dari ketiga kolom, yaitu `Age`, `AnnualIncome`, dan `FamilyMembers` namun tidak begitu signifikan

```
[ ] df[cat].describe()
```

	Employment Type	GraduateOrNot	FrequentFlyer	EverTravelledAbroad	TravelInsurance	ChronicDiseases
count	1249	1249	1249	1249	1249	1249
unique	2	2	2	2	2	2
top	Private Sector/Self Employed	Yes	No	No	No	No
freq	876	1047	954	1005	766	833

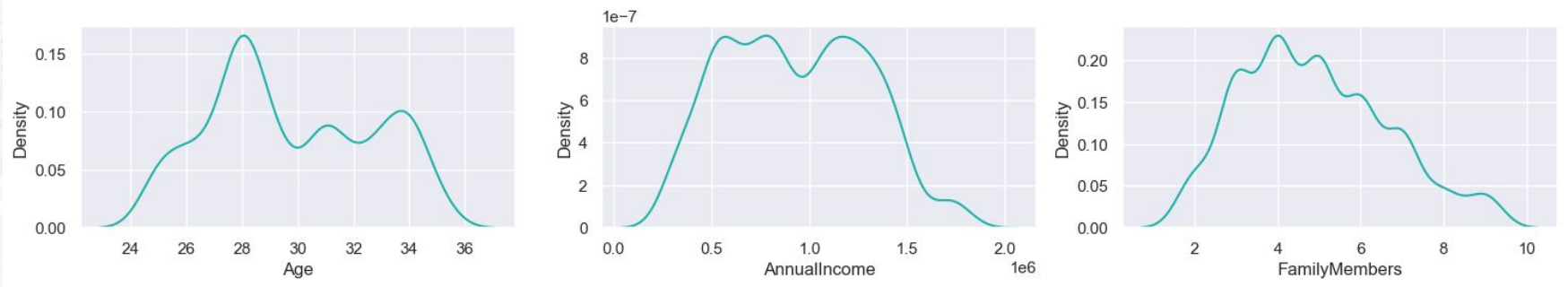
EDA, Insights & Visualization

2.) Univariate Analysis (1/3)



Feature Numerical kami menggunakan boxplot dan displot. Berikut hasil pengamatan kami:

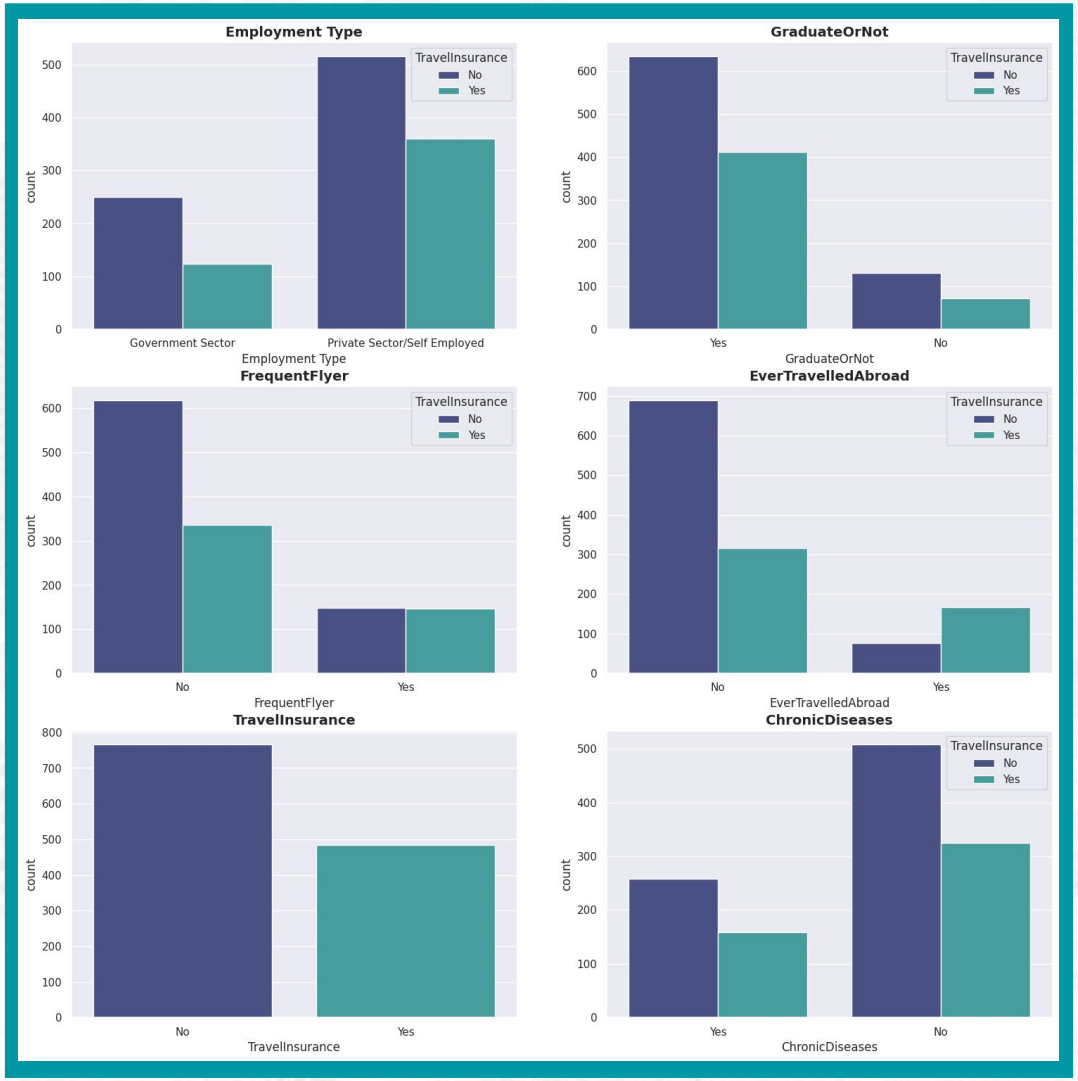
- Dapat dilihat dari boxplot disamping, ketiga kolom (`Age`, `AnnualIncome`, dan `FamilyMembers`) tidak terdapat outliers
- Pada displot dapat dilihat kolom `Age`, `AnnualIncome`, dan `FamilyMembers` menunjukkan distribusi hampir normal, tidak ada skewness.



**Univariate Analysis : Cara untuk memahami satu variabel pada satu waktu. Analisis ini merupakan cara awal yang baik untuk menggali data dan mengidentifikasi pola atau abnormalitas*

EDA, Insights & Visualization

2.) Univariate Analysis (2/3)

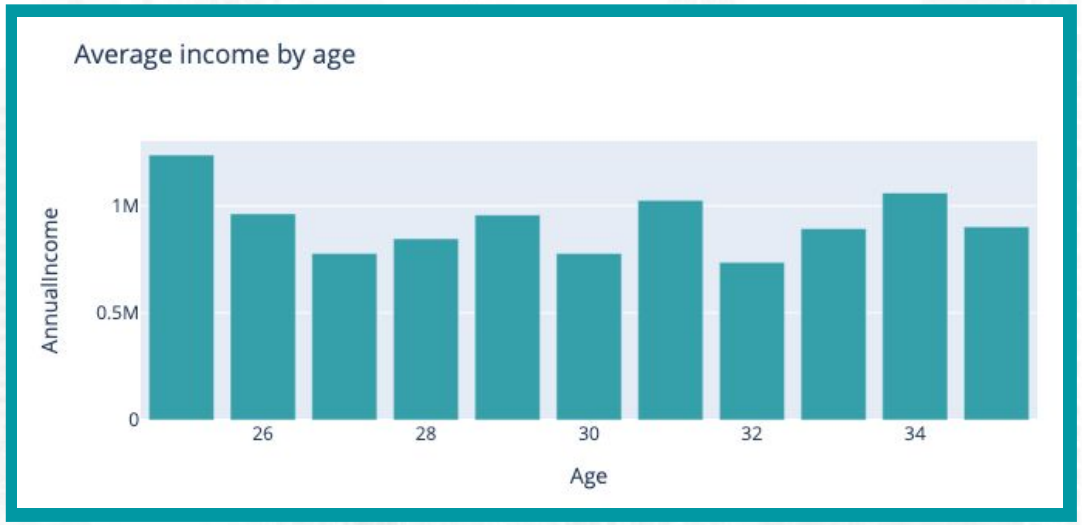


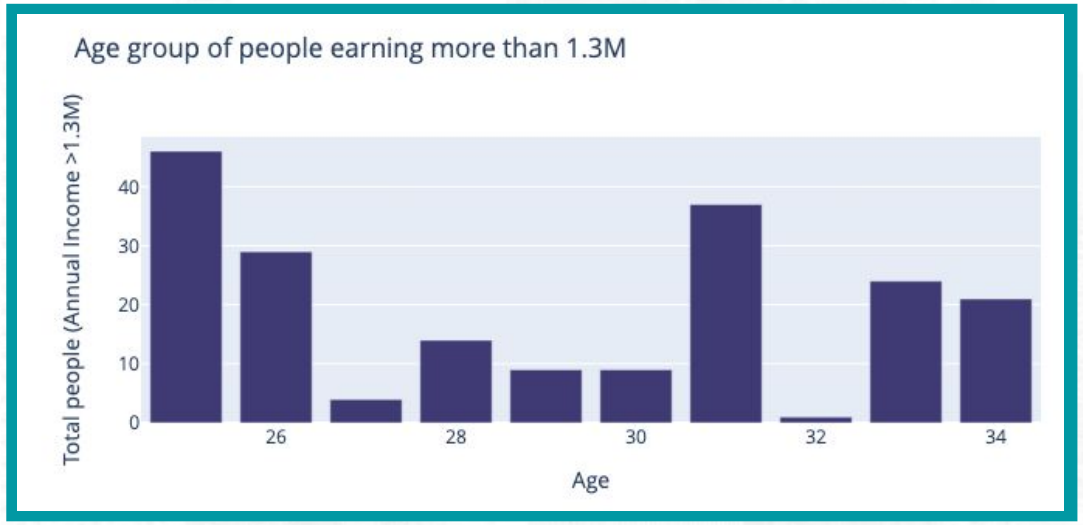
Feature Categorical kami coba untuk melihat pengaruh masing-masing feature terhadap pembelian Travel Insurance, berikut hasil interpretasinya:

- Baik customer yang bekerja di **Pemerintahan** maupun **Swasta** cenderung **tidak membeli Asuransi perjalanan**. Namun rasio pembelian Asuransi di **Sektor Swasta lebih tinggi** (876 orang) dibandingkan **Pemerintah** (373 orang).
- **Tidak terdapat perbedaan** yang signifikan dalam keputusan pembelian **Asuransi Perjalanan** antara customer yang tamat **sarjana atau** yang **tidak**.
- Customer yang mengidap **Penyakit Kronis** atau **tidak** tampaknya **tidak memiliki pengaruh** yang signifikan terhadap kecenderungan untuk **membeli Asuransi**.
- **Not Frequent Flyer**, memiliki **potensi yang lebih tinggi** untuk membeli **Asuransi.**
- Pelanggan yang **pernah bepergian ke Luar Negeri cenderung membeli Asuransi.**

EDA, Insights & Visualization

2.) Univariate Analysis (3/3)





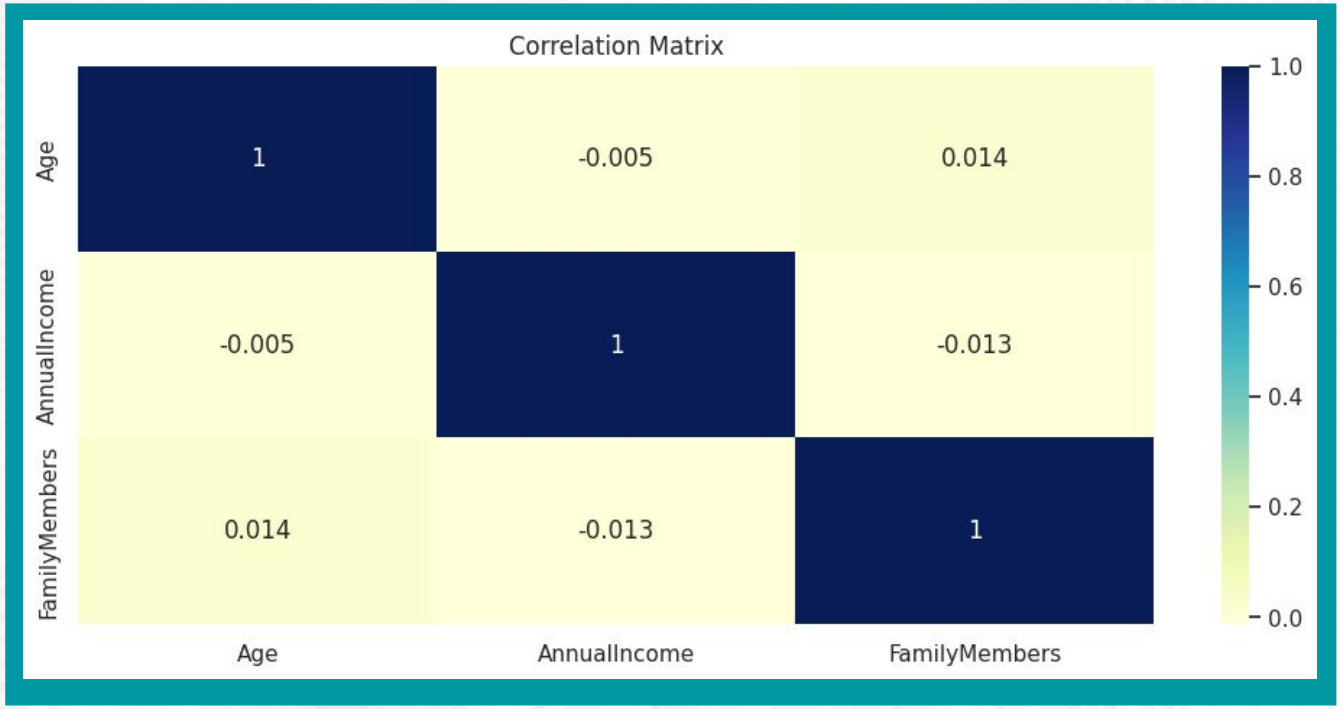
Rata-rata **income tertinggi** berada di customer dengan umur **25 tahun**

Dimana rata-rata tersebut menunjukkan income sebesar lebih dari 1.3M

EDA, Insights & Visualization

3.) Multivariate Analysis (1/3)

Analisis Multivariat merupakan metode statistik yang memungkinkan melakukan penelitian terhadap satu atau lebih dari dua variabel secara bersamaan. Untuk melihat korelasi feature kami menggunakan heatmap.



Dari correlation heatmap dapat dilihat bahwa:

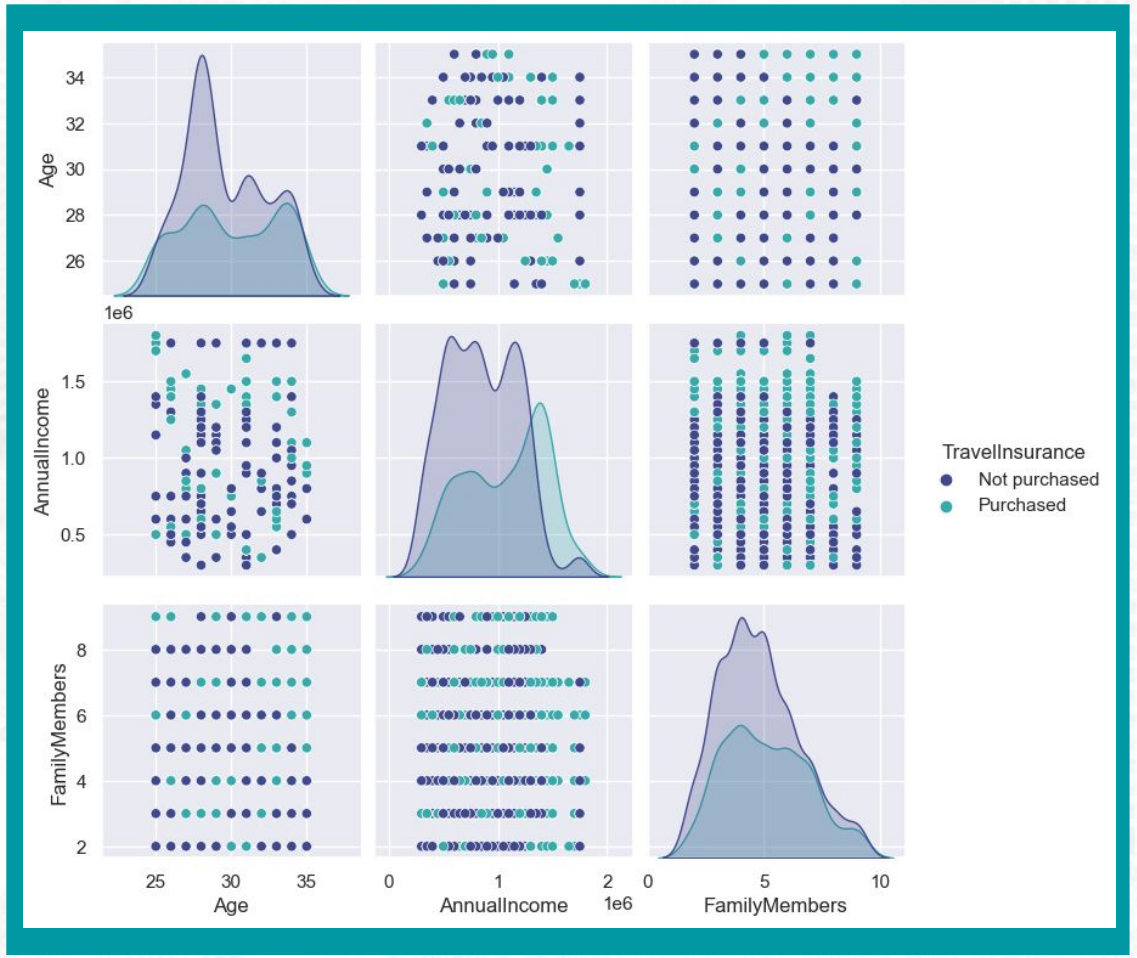
- **Tidak ada** feature numeric yang memiliki **korelasi yang kuat** satu sama lain.
- Annual Income dan Family Members memiliki korelasi negatif yang sangat rendah.
- Annual Income dan Umur memiliki korelasi negatif yang rendah.

**Multivariate Analysis : Cara kita melihat dua atau lebih variabel pada saat yang sama untuk memahami bagaimana mereka berkaitan dan berinteraksi.*

EDA, Insights & Visualization

3.) Multivariate Analysis (2/3)

Untuk melihat distribusi dan korelasi feature kami menggunakan pairplot.



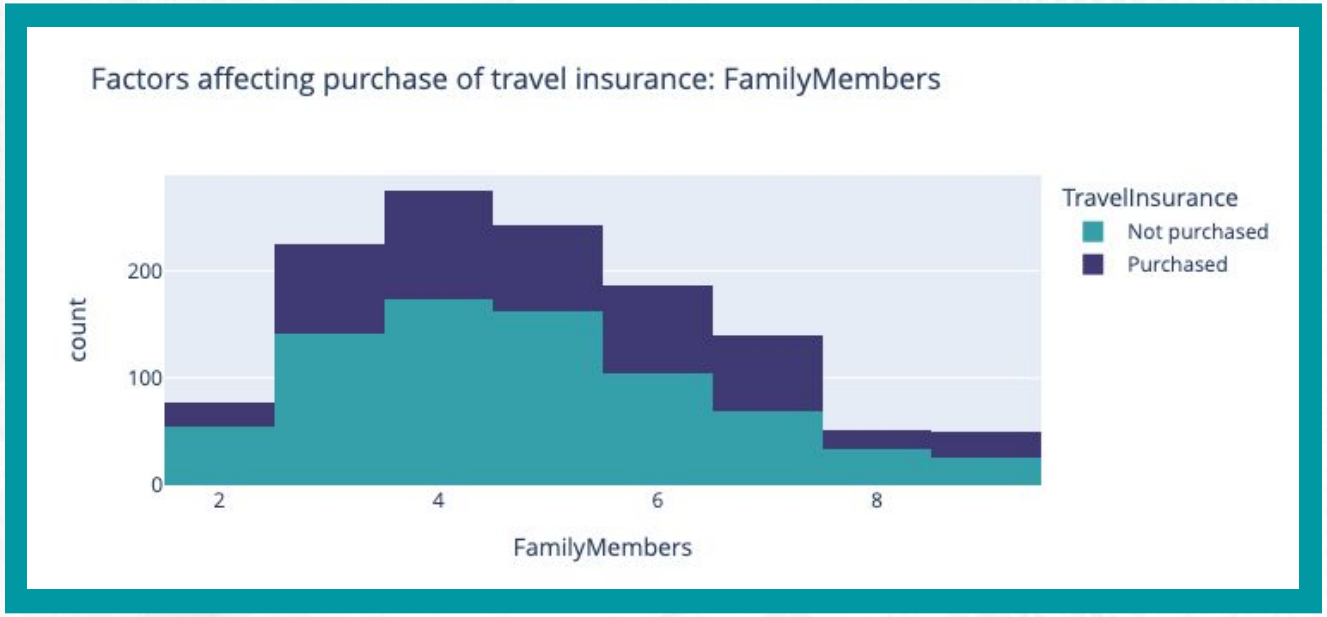
Dari pairplot di samping dapat dilihat bahwa:

- Terdapat **penurunan jumlah pelanggan** di kedua tipe sektor pekerjaan pada kondisi tertentu seperti age, annual income dan jumlah family member. Kita bisa melihat lebih detail mengenai hal ini

EDA, Insights & Visualization

3.) Multivariate Analysis (3/3)

Kami pun melihat apakah faktor dari banyaknya Family Members mempengaruhi pembelian Asuransi Perjalanan, dan didaparkan data berikut:



- Jumlah Family Member yang beranggotakan **4 orang** mendominasi sebagai group yang membeli Asuransi Perjalanan dibandingkan dengan jumlah anggota yang lain.
- Semakin banyak anggota keluarga (>4 orang), peminat Asuransi Perjalanan **semakin menurun**.

EDA, Insights & Visualization

4.) Business Insight

- Perusahaan dapat mempertimbangkan untuk mengarahkan strategi pemasaran asuransi perjalanan lebih khusus ke pelanggan di sektor swasta, mengingat rasio pembelian yang lebih tinggi di sektor ini.
- Perusahaan tidak perlu mempertimbangkan lebih jauh mengenai tingkat pendidikan, karena tidak ada perbedaan yang signifikan terhadap keputusan pembelian.
- Meskipun penyakit kronis tidak tampak mempengaruhi keputusan pembelian, perusahaan dapat mempertimbangkan penawaran atau manfaat asuransi yang lebih menarik bagi pelanggan dengan kondisi kesehatan tertentu.
- Perusahaan dapat mempertimbangkan strategi pemasaran yang menawarkan berbagai promo untuk paket wisata dengan menargetkan customer yang tidak sering berpergian dan yang belum pernah bepergian ke luar negeri.
- Perusahaan dapat menawarkan paket asuransi keluarga khususnya untuk keluarga beranggotakan 4 orang.
- Perusahaan dapat menawarkan paket member travel insurance dengan segmen annual income.

Git

Website URL : https://github.com/BYTE-BLAZERS/Final-Project-Byte-Blazers.git
GITHUB CLI URL : gh repo clone BYTE-BLAZERS/Final-Project-Byte-Blazers