

2022

Car Insurance



Omar Ali
Epsilon
1/1/2022

About Dataset:

The company has shared its annual car insurance data. Now, you must find out the real customer behaviors over the data. This is customer data with their vehicle insurance policies. Details about customers and the insurance taken for their vehicles are provided which can be explored to segment similar kinds of customers.

From Kaggle:

<https://www.kaggle.com/datasets/sagnik1511/car-insurance-data>

Data Contains:

10000 Rows

19 Columns

Features:

- 1- ID: To Identify the Persons**
- 2- AGE: From 16 To 65+**
- 3- GENDER: Male, Female**
- 4- RACE: The majority**
- 5- DRIVING_EXPERINCE: From 0 To 30+**
- 6- EDUCATION: The Education a Person has Received**
- 7- INCOME**
- 8- CREDIT_SCORE: a Number Rate a Consumers Creditworthiness**
- 9- VEHICLE_OWNERSHIP**
- 10- VEHICLE_YEAR**
- 11- MARRIED**
- 12- CHILDREN**
- 13- POSTAL CODE**
- 14- ANNUAL_MILEAGE: The Number of Miles You Drive Each Year**
- 15- VEHICLE_TYPE**
- 16- SPEEDING_VIOLATIONS**
- 17- DUIS**
- 18- PAST_ACCIDENTS**
- 19- OUTCOME**

Question Analysis:

- 1- Is There a relationship between driving_experience and outcome**
- 2- What is the maximum outcome per gender**
- 3- Is there a relationship between education and outcome**
- 4- Is there a relationship between income and outcome**
- 5- Is there a relationship between age and outcome**
- 6- What is the maximum outcome per age**

Information about data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 19 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   ID                                     10000 non-null  int64
1   AGE                                   10000 non-null  object
2   GENDER                               10000 non-null  object
3   RACE                                 10000 non-null  object
4   DRIVING_EXPERIENCE                   10000 non-null  object
5   EDUCATION                           10000 non-null  object
6   INCOME                               10000 non-null  object
7   CREDIT_SCORE                         9018 non-null   float64
8   VEHICLE_OWNERSHIP                   10000 non-null  float64
9   VEHICLE_YEAR                        10000 non-null  object
10  MARRIED                             10000 non-null  float64
11  CHILDREN                            10000 non-null  float64
12  POSTAL_CODE                         10000 non-null  int64
13  ANNUAL_MILEAGE                      9043 non-null   float64
14  VEHICLE_TYPE                        10000 non-null  object
15  SPEEDING_VIOLATIONS                 10000 non-null  int64
16  DUIS                                10000 non-null  int64
17  PAST_ACCIDENTS                      10000 non-null  int64
18  OUTCOME                             10000 non-null  float64
dtypes: float64(6), int64(5), object(8)
```

	ID	CREDIT_SCORE	VEHICLE_OWNERSHIP	MARRIED	CHILDREN	POSTAL_CODE	ANNUAL_MILEAGE	SPEEDING_VIOLATIONS	DUIS	PAST_ACCIDENTS	OUTCOME
count	1000	9018.000000	10000.000000	1000.000000	1000.000000	1000.000000	9043.000000	10000.000000	10000.000000	10000.000000	1000.000000
mean	5005.21906800	0.515813	0.697000	0.498200	0.688800	19864.548400	11697.003207	1.482900	0.23920	1.056300	0.313300
std	290030.768758	0.137688	0.459578	0.500022	0.463008	18915.613855	2818.434528	2.241966	0.55499	1.652454	0.463858

	ID	CREDIT_SCORE	VEHICLE_OWNERSHIP	MARRIED	CHILDREN	POSTAL_CODE	ANNUAL_MILEAGE	SPEEDING_VIOLATIONS	DUIS	PAST_ACCIDENTS	OUTCOME
min	101.00000	0.053358	0.000000	0.00000	0.00000	10238.00000	2000.00000	0.000000	0.00000	0.00000	0.00000
25%	249638.50000	0.417191	0.000000	0.00000	0.00000	10238.00000	10000.00000	0.000000	0.00000	0.00000	0.00000
50%	501777.00000	0.525033	1.000000	0.00000	1.00000	10238.00000	12000.00000	0.000000	0.00000	0.00000	0.00000
75%	753974.50000	0.618312	1.000000	1.00000	1.00000	32765.00000	14000.00000	2.000000	0.00000	2.00000	1.00000
max	999976.00000	0.960819	1.000000	1.00000	1.00000	92101.00000	22000.00000	22.000000	6.00000	15.00000	1.00000

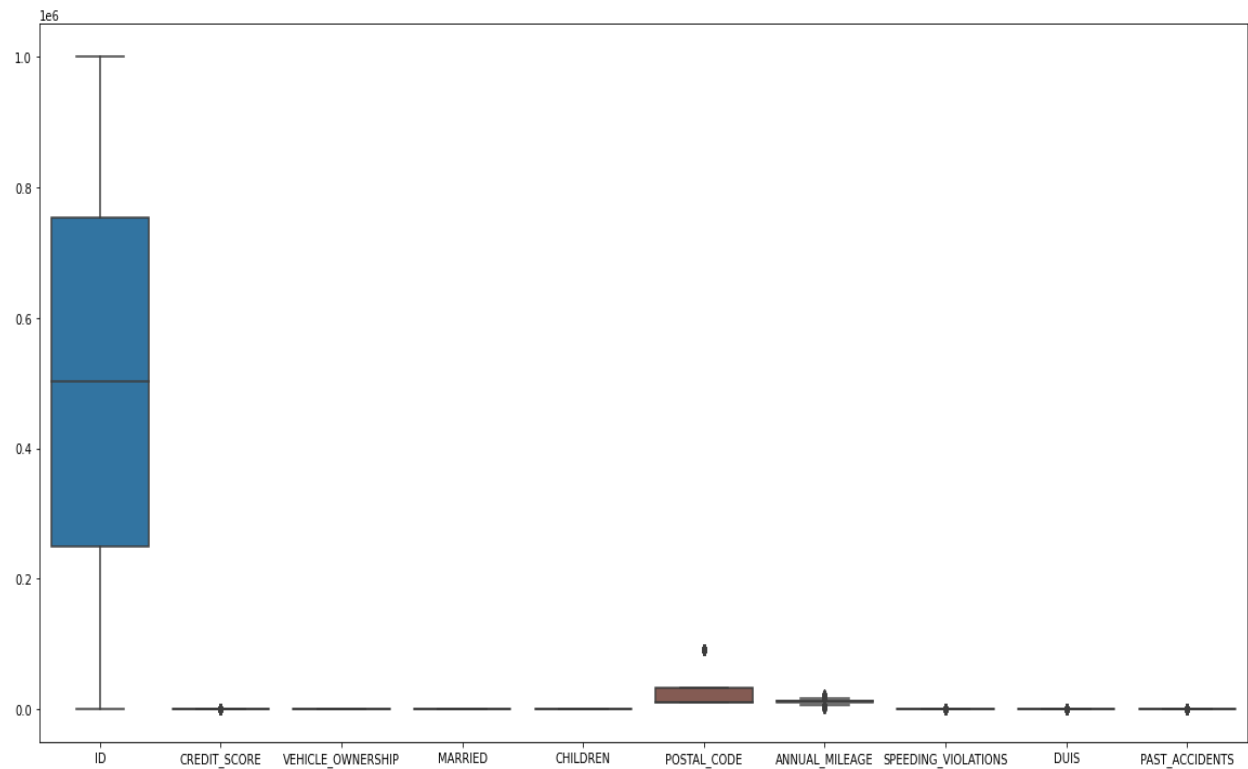
Missing value:

```
CREDIT_SCORE    982  ANNUAL_MILEAGE    957  ID      0  MARRIED    0
PAST_ACCIDENTS  0  DUIS      0  SPEEDING_VIOLATIONS  0  VEHICLE_TYPE  0
POSTAL_CODE     0  CHILDREN    0  VEHICLE_YEAR    0  AGE      0
VEHICLE_OWNERSHIP  0  INCOME    0  EDUCATION    0  DRIVING_EXPERIENCE
0  RACE    0  GENDER    0  OUTCOME    0
```

after processing:

```
ID 0 AGE 0 GENDER 0 RACE 0 DRIVING_EXPERIENCE 0 EDUCATION 0
INCOME 0 CREDIT_SCORE 0 VEHICLE_OWNERSHIP 0 VEHICLE_YEAR 0
MARRIED 0 CHILDREN 0 POSTAL_CODE 0 ANNUAL_MILEAGE 0
VEHICLE_TYPE 0 SPEEDING_VIOLATIONS 0 DUIS 0 PAST_ACCIDENTS 0
OUTCOME 0
```

Outliers:



The number of outliers are : 2955

After processing:

0	ID	10000	non-null	int64
1	AGE	10000	non-null	object
2	GENDER	10000	non-null	object
3	RACE	10000	non-null	object
4	DRIVING_EXPERIENCE	10000	non-null	object
5	EDUCATION	10000	non-null	object
6	INCOME	10000	non-null	object
7	CREDIT_SCORE	10000	non-null	float64
8	VEHICLE_OWNERSHIP	10000	non-null	float64
9	VEHICLE_YEAR	10000	non-null	object
10	MARRIED	10000	non-null	float64
11	CHILDREN	10000	non-null	float64
12	POSTAL_CODE	10000	non-null	int64
13	ANNUAL_MILEAGE	10000	non-null	float64
14	VEHICLE_TYPE	10000	non-null	object
15	SPEEDING_VIOLATIONS	10000	non-null	int64
16	DUI	10000	non-null	int64
17	PAST_ACCIDENTS	10000	non-null	int64
18	OUTCOME	10000	non-null	float64

there are no outliers

question analysis:

Relation Between Age & Outcome:

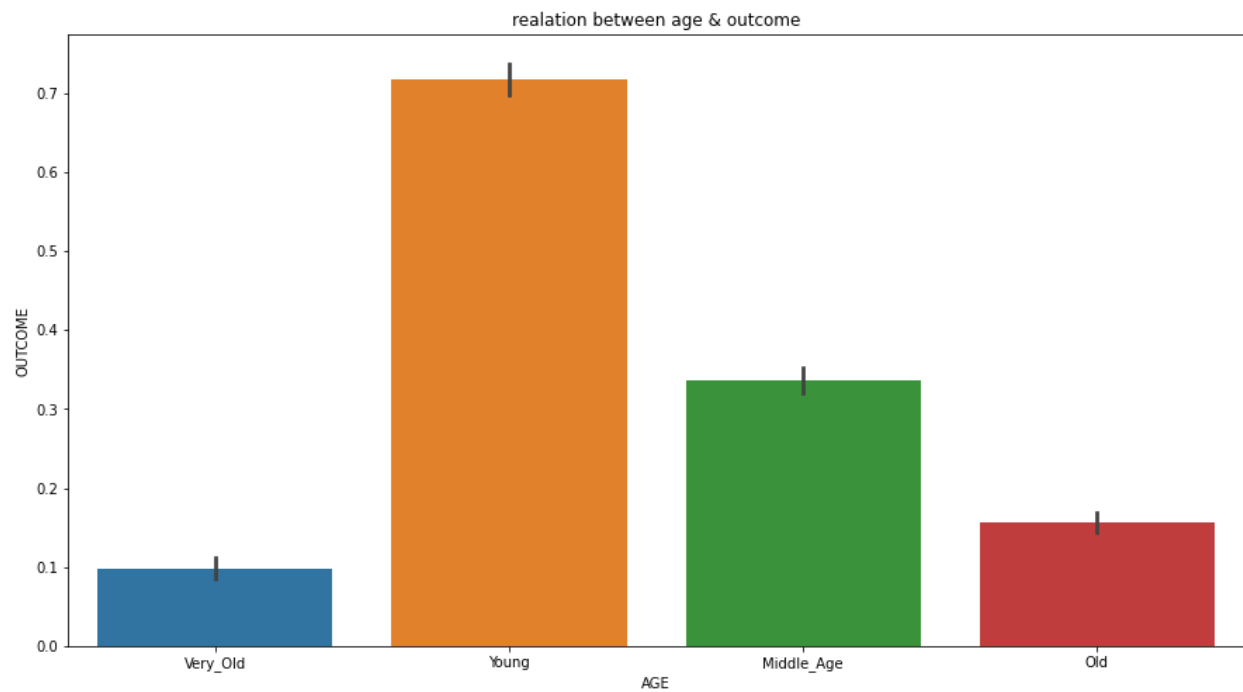
OUTCOME, AGE

Young 1448.0

Middle_Age 1032.0

Old 457.0

Very Old 196.0



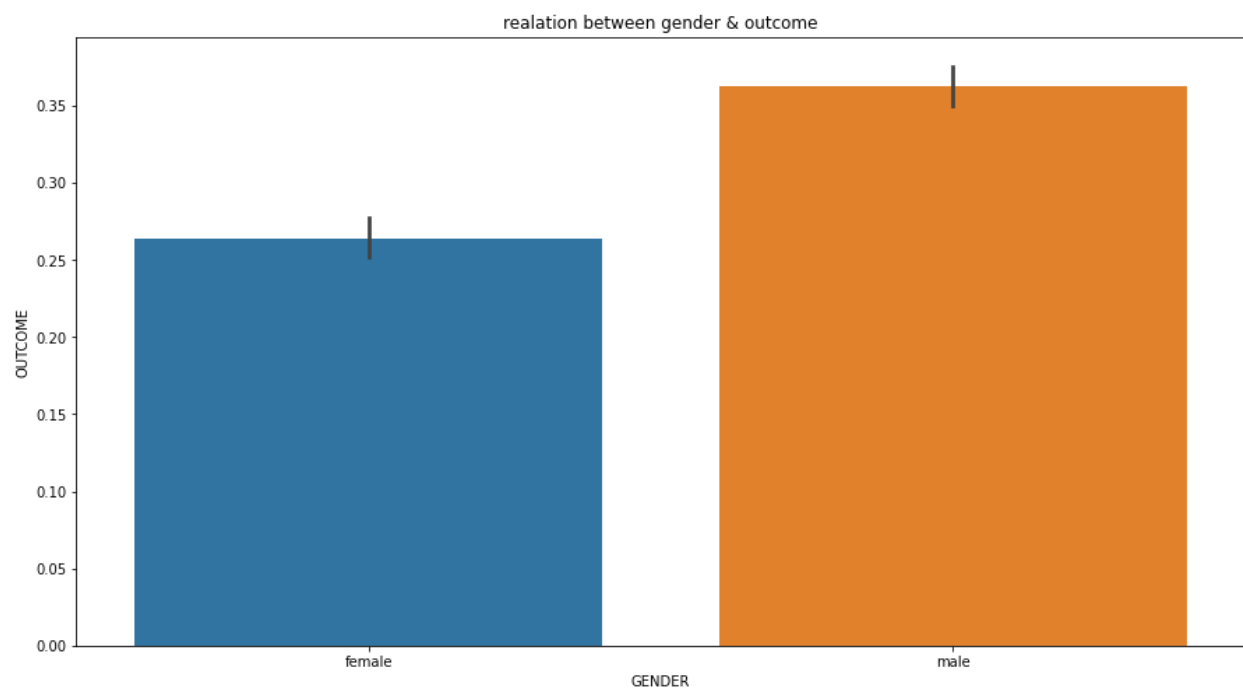
relation between gender & outcome:

OUTCOME

GENDER

male 1812.0

female 1321.0

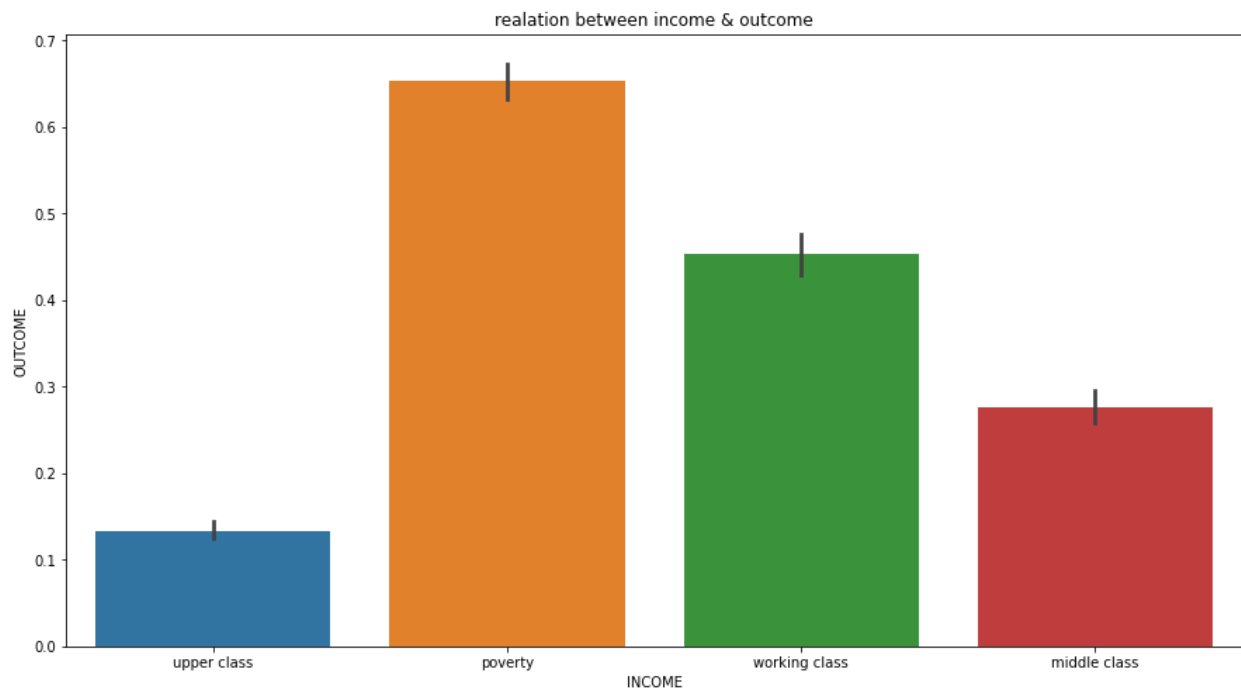


relation between income & outcome:

OUTCOME

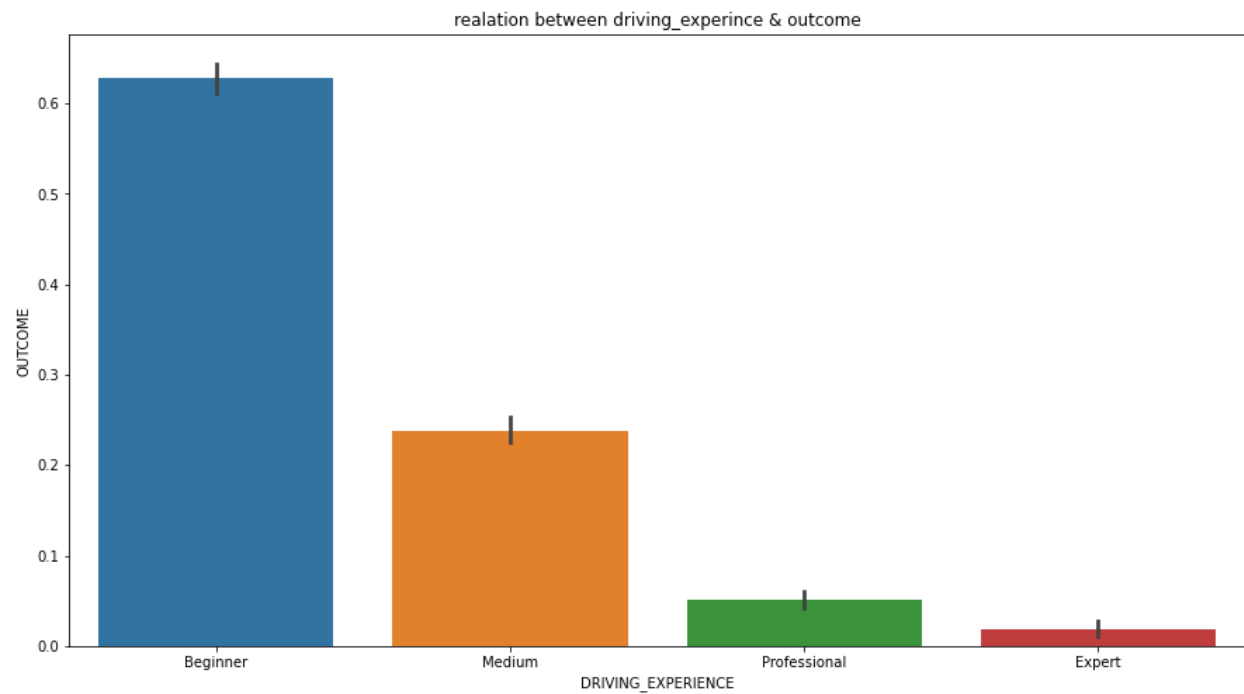
INCOME

poverty	1186.0
working class	776.0
middle class	592.0
upper class	579.0



relation between driving_experience & outcome:

OUTCOME	
DRIVING_EXPERIENCE	
Beginner	2217.0
Medium	787.0
Professional	109.0
Expert	20.0



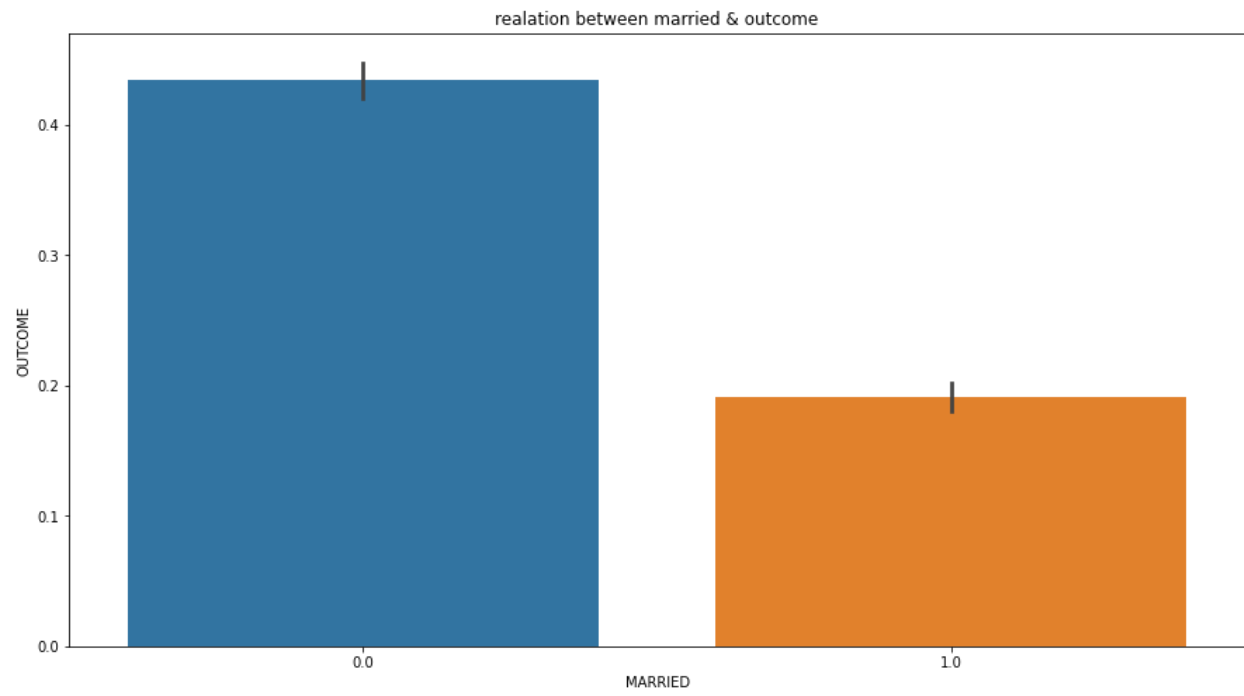
relation between married & outcome:

OUTCOME

MARRIED

0.0 2180.0

1.0 953.0



Preprocessing steps:

Handle Missing Values

- Handling Categorical Data
- Handling Outliers
- Split Data into Train and Test
- Feature Scaling