Understanding the data

```
#list of columns in our dataset
col_list = df.columns
print(col_list)
```

Glimpse Of Data

Top 5 entries of the dataset
df.head()

months_as_customer	age	policy_number	<pre>policy_bind_date</pre>	<pre>policy_state</pre>	policy_cs
5	37	939011	16-07-2002	IN	250/50

Bottom 5 entries of the dataset
df.tail()

	months_as_customer	age	policy_number	<pre>policy_bind_date</pre>	policy_state	polic
10206	91	16	524932	23-02-2014	IN	10
10207	266	29	128125	12-10-2006	ОН	25
10208	332	40	494839	21-03-2008	ОН	1(
10209	316	48	122384	28-02-2004	IL	500
10210	113	20	614699	05-10-1999	ОН	1(

df.describe()

	months_as_customer	age	policy_number	policy_deductable	policy_anr
count	10211.000000	10211.000000	1.021100e+04	10211.000000	
mean	213.467927	39.050142	5.474680e+05	1159.044168	
std	133.639732	11.508964	3.034069e+05	621.773731	
min	0.000000	2.000000	4.410000e+02	500.000000	
25%	106.000000	31.000000	3.095050e+05	500.000000	
50%	202.000000	38.000000	5.364750e+05	1000.000000	
75%	303.000000	47.000000	7.717955e+05	2000.000000	
max	747.000000	79.000000	1.615353e+06	2000.000000	

#unique entries in each column of dataframe
df.nunique()

months_as_customer	614
age	76
policy_number	10163
policy_bind_date	951
policy_state	3
policy_csl	3
policy_deductable	3
policy_annual_premium	9706
umbrella_limit	11

10045

2

7

14

insured_zip
insured_sex

insured_education level

insured occupation

```
insured_hobbies
                                       20
     insured relationship
                                        6
     capital.gains
                                     4940
     capital.loss
                                     4871
     incident date
                                       60
                                        4
     incident_type
     collision_type
                                        4
     incident_severity
                                        4
     authorities_contacted
                                        5
     incident state
                                        7
                                        7
     incident city
     incident_location
                                     1000
     incident_hour_of_the_day
                                       25
     number_of_vehicles_involved
                                        6
     property_damage
                                        3
     bodily injuries
                                        5
    witnesses
                                        7
     police_report_available
                                        3
    total_claim_amount
                                     4816
     injury_claim
                                     1825
     property_claim
                                     1858
                                     4147
    vehicle_claim
    auto_make
                                       14
     auto_model
                                       39
                                       35
     auto year
                                        2
     fraud_reported
     dtype: int64
cols = ['policy_state','policy_csl','policy_deductable','insured_sex','insured_education_l
        'incident_type','collision_type','incident_severity','authorities_contacted','prop
        'police_report_available', 'auto_make']
for i in cols:
 temp = df[i].unique()
 print(i,":",temp)
     policy_state : ['IN' 'IL' 'OH']
     policy_csl : ['250/500' '500/1000' '100/300']
     policy deductable : [ 500 1000 2000]
     insured_sex : ['FEMALE' 'MALE']
     insured_education_level : ['Associate' 'MD' 'High School' 'PhD' 'JD' 'Masters' 'Colle
     insured_occupation : ['priv-house-serv' 'exec-managerial' 'farming-fishing' 'transpor
      'armed-forces' 'tech-support' 'protective-serv' 'prof-specialty'
      'machine-op-inspct' 'other-service' 'adm-clerical' 'handlers-cleaners'
      'craft-repair' 'sales']
     incident_type : ['Single Vehicle Collision' 'Multi-vehicle Collision' 'Parked Car'
      'Vehicle Theft']
     collision type : ['Front Collision' 'Rear Collision' '?' 'Side Collision']
     incident_severity : ['Minor Damage' 'Total Loss' 'Trivial Damage' 'Major Damage']
     authorities contacted : ['Other' 'None' 'Police' 'Fire' 'Ambulance']
     property_damage : ['?' 'NO' 'YES']
     police report available : ['YES' '?' 'NO']
     auto_make : ['Saab' 'Chevrolet' 'Toyota' 'Honda' 'Accura' 'BMW' 'Audi' 'Suburu'
      'Dodge' 'Nissan' 'Jeep' 'Ford' 'Mercedes' 'Volkswagen']
```

- Cleaning Data

```
df = df.replace(to_replace="Y",value="YES")
df = df.replace(to_replace="N",value="NO")
#df = df.drop(df[df.score < 50].index)</pre>
df = df.drop(['insured_hobbies'],axis=1)
#df.isnull().sum()
#looking for outliers
df['age'].describe()
     count
              10211.000000
     mean
                 39.050142
     std
                 11.508964
                  2.000000
     min
     25%
                 31.000000
     50%
                 38.000000
     75%
                 47.000000
                 79.000000
     max
     Name: age, dtype: float64
corelation = df.corr()
sns.set(rc = {'figure.figsize':(19,10)})
sns.heatmap(corelation,xticklabels=corelation.columns,
            yticklabels=corelation.columns,annot=True)
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f573f705e10>

months_as_customer	1	0.64	0.015	0.02	-4.7e-05	-0.0038	-0.0037	0.0085	-0.0022	0.046	-0.0065	0.0018	0.034	0.035	0.033	0.015	0.058	-0.011		1.0
age	0.64	1	0.02	0.031	0.015	-0.019	0.0043	0.019	-0.0043	0.054	-0.0062	0.00016	0.03	0.032	0.039	0.032	0.037	-0.018		
policy_number	0.015	0.02	1	0.011	0.018	-0.0011	0.016	0.0051	-0.028	-0.0029	0.033	-0.007	0.017	-0.012	-0.014	0.001	-0.011	-0.0077		
policy_deductable	0.02	0.031	0.011	1	-0.027	-0.0089	0.023	0.01	-0.0062	0.023	0.049	-0.029	0.026	-0.011	0.015	0.016	-0.016	0.022		- 0.8
policy_annual_premium	-4.7e-05	0.015	0.018	-0.027	1	0.0072	0.018	0.017	-0.0075	-0.015	-0.046	0.0067	0.0067	-0.0026	-0.03	-0.0067	0.01	-0.034		
umbrella_limit	-0.0038	-0.019	-0.0011	-0.0089	0.0072	1	0.038	-0.023	0.015	-0.021	-0.019	0.033	-0.018	-0.016	0.0026	-0.0089	-0.03	0.016		
insured_zip	-0.0037	0.0043	0.016	0.023	0.018	0.038	1	-0.0066	-0.021	-0.034	0.023	0.027	0.017	-0.017	-0.012	0.022	-0.021	-0.0089		- 0.6
capital.gains	0.0085	0.019	0.0051	0.01	0.017	-0.023	-0.0066	1	0.021	0.014	0.02	0.0056	-0.0039	-0.0066	0.00069	9-0.015	-0.0037	-0.0035		- 0.6
capital.loss	-0.0022	-0.0043	-0.028	-0.0062	-0.0075	0.015	-0.021	0.021	1	0.015	0.0015	0.037	0.0061	0.01	0.026	0.00084	0.012	0.0097		
incident_hour_of_the_day	0.046	0.054	-0.0029	0.023	-0.015	-0.021	-0.034	0.014	0.015	1	0.091	-0.0013	0.00036	0.11	0.091	0.11	0.11	0.025		
number_of_vehicles_involved	-0.0065	-0.0062	0.033	0.049	-0.046	-0.019	0.023	0.02	0.0015	0.091	1	-0.005	-0.0076	0.11	0.095	0.089	0.11	0.029		- 0.4
bodily_injuries	0.0018	0.00016	-0.007	-0.029	0.0067	0.033	0.027	0.0056	0.037	-0.0013	-0.005	1	0.0058	0.045	0.054	0.045	0.039	0.017		
witnesses	0.034	0.03	0.017	0.026	0.0067	-0.018	0.017	-0.0039	0.0061-	0.00036	5-0.0076	0.0058	1	0.02	-0.013	0.062	-0.0093	0.069		
total_claim_amount	0.035	0.032	-0.012	-0.011	-0.0026	-0.016	-0.017	-0.0066	0.01	0.11	0.11	0.045	0.02	1	0.5	0.51	0.63	-0.0068		- 0.2
injury_claim	0.033	0.039	-0.014	0.015	-0.03	0.0026	-0.012	-0.00069	0.026	0.091	0.095	0.054	-0.013	0.5	1	0.3	0.42	0.016		0.2
property_claim	0.015	0.032	0.001	0.016	-0.0067	-0.0089	0.022	-0.015	0.00084	0.11	0.089	0.045	0.062	0.51	0.3	1	0.44	0.011		
vehicle_claim	0.058	0.037	-0.011	-0.016	0.01	-0.03	-0.021	-0.0037	0.012	0.11	0.11	0.039	-0.0093		0.42	0.44	1	-0.019		
auto_year	-0.011	-0.018	-0.0077	0.022	-0.034	0.016	-0.0089	-0.0035	0.0097	0.025	0.029	0.017	0.069	-0.0068	0.016	0.011	-0.019	1		- 0.0
	er	age	er	e e	E	nit	di	SU	SS	ay	pa	es	es	nt	E	E	Ē	je je	_	
	customer	ю	policy_number	oolicy_deductable	annual_premium	.mbrella_limit	insured_zip	apital.gains	capital.loss	the_d	volv	oodily_injuries	witnesses	claim_amount	injury_claim	property_claim	e_claim	auto_year		
	as cu		olicy_	y_dec	ual_p	mbre	inst	capit	ğ	I of	cles	odilly	×	ia M	inju	opert	vehicle	æ		
	months		ā	polic		_				cident_hour_of_the_day	er_of_vehicles_involved	q		botal_c		ā				
	Ĕ				policy					cider	er_of			4						



-1.0