Predicting Credit Risk

Caleb Puckett

cpuckett@bellarmine.edu

February 6, 2024

The data used in this exploratory data analysis can be found at the following link: https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction/data. I chose this dataset because I am interested in the topic it discusses. It also seems well prepared for analysis. It was found on the website "Kaggle" and there is a description of each variable made available to viewers.

The dataset contains many variables; these can be seen in the table below. The variables which need further explanation are income type, marital status, housing type, days employed, mobile phone, work phone, phone, months balance, and status. Income type refers to the type of income received, with the most prominent type being "working." Other categories include commercial associate, pensioner, state servant, and student. This list seems rather vague, so it may not provide very meaningful results. Marital status includes civil marriage, separated, single / not married, widow, and married, with married being the most prominent value. Housing type includes outputs of co-op apartment, house / apartment, municipal apartment, office apartment, rented apartment, and with parents. The most prominent of these outputs is house / apartment, although the vague and similar outputs may prevent this information from being especially useful.

Days employed shows the number of days since the start of employment for a given person. The range of this value is shown in the table below, where negative numbers are typical. The positive numbers which are returned convey that the specified person is unemployed. There is no reason for this positive number to be so large, but this does not mean anything of substance.

Mobile phone, work phone, and phone are all dummy variables which return a one if the given person owns the specified item, and a zero if they do not. Phone does not appear clear or consistent, so it may need to be removed after further analysis. Months balance reflects the month of the taken record. Negative numbers mean the data point was taken from previous months, with the attached number specifying the number of months. Status is the response variable, which tells us the status of the payment. C means that the debt was paid off that month, X means that there was no loan for the month, and the numbers returned tell how late the payment is.

Name	Data Type	Range of Values	Percentage of Missing Data
ID	Nominal	0 – 5150337	0%
CHILDREN	Interval	0 – 19	0%
INCOME	Interval	27,000 – 1,575,000	0%
INCOME_TYPE	Nominal	5 String Values	0%
EDUCATION	Nominal	5 String Values	0%
MARITAL_STATUS	Nominal	5 String Values	0%
HOUSING_TYPE	Nominal	6 String Values	0%
DAYS_SINCE_BIRTH	Ordinal	(-25152) – (-7489)	0%
DAYS_EMPLOYED	Ordinal	(-15713) – 365243	0%
MOBILE_PHONE	Nominal	0 – 1	0%
WORK_PHONE	Nominal	0 – 1	0%
PHONE	Nominal	0 – 1	0%
EMAIL	Nominal	0 – 1	0%
OCCUPATION_TYPE	Nominal	18 String Values	30.87%
FAM_SIZE	Interval	1 – 20	0%
MONTHS_BALANCE	Ordinal	(-60) – 0	0%
GENDER_F	Nominal	0 – 1	0%
GENDER_M	Nominal	0 – 1	0%
CAR_N	Nominal	0 – 1	0%
CAR_Y	Nominal	0 – 1	0%
REALTY_N	Nominal	0 – 1	0%
REALTY_Y	Nominal	0 – 1	0%
STATUS	Ordinal	8 String Values	0%

Because there are often multiple rows with the same ID values, which show consecutive months of the same individual's payment history, there are considerably more rows than there are IDs. There were originally two datasets: one which contained ID, status, and months balance, and another which contained all of the other variables, including a repeat of ID. These two datasets were combined into one dataset with all information and where the IDs of both were equal.

	ID	CHILDREN	INCOME	DAYS_SINCE_BIRTH	DAYS_EMPLOYED	MOBILE_PHONE	WORK_PHONE	PHONE
count	7.777150e+05	777715.000000	7.777150e+05	777715.000000	777715.000000	777715.0	777715.000000	777715.000000
mean	5.078743e+06	0.428082	1.885348e+05	-16124.937046	57775.825016	1.0	0.231818	0.300965
std	4.180442e+04	0.745755	1.016225e+05	4104.304018	136471.735391	0.0	0.421993	0.458678
min	5.008804e+06	0.000000	2.700000e+04	-25152.000000	-15713.000000	1.0	0.000000	0.000000
25%	5.044568e+06	0.000000	1.215000e+05	-19453.000000	-3292.000000	1.0	0.000000	0.000000
50%	5.069530e+06	0.000000	1.620000e+05	-15760.000000	-1682.000000	1.0	0.000000	0.000000
75%	5.115551e+06	1.000000	2.250000e+05	-12716.000000	-431.000000	1.0	0.000000	1.000000
max	5.150487e+06	19.000000	1.575000e+06	-7489.000000	365243.000000	1.0	1.000000	1.000000

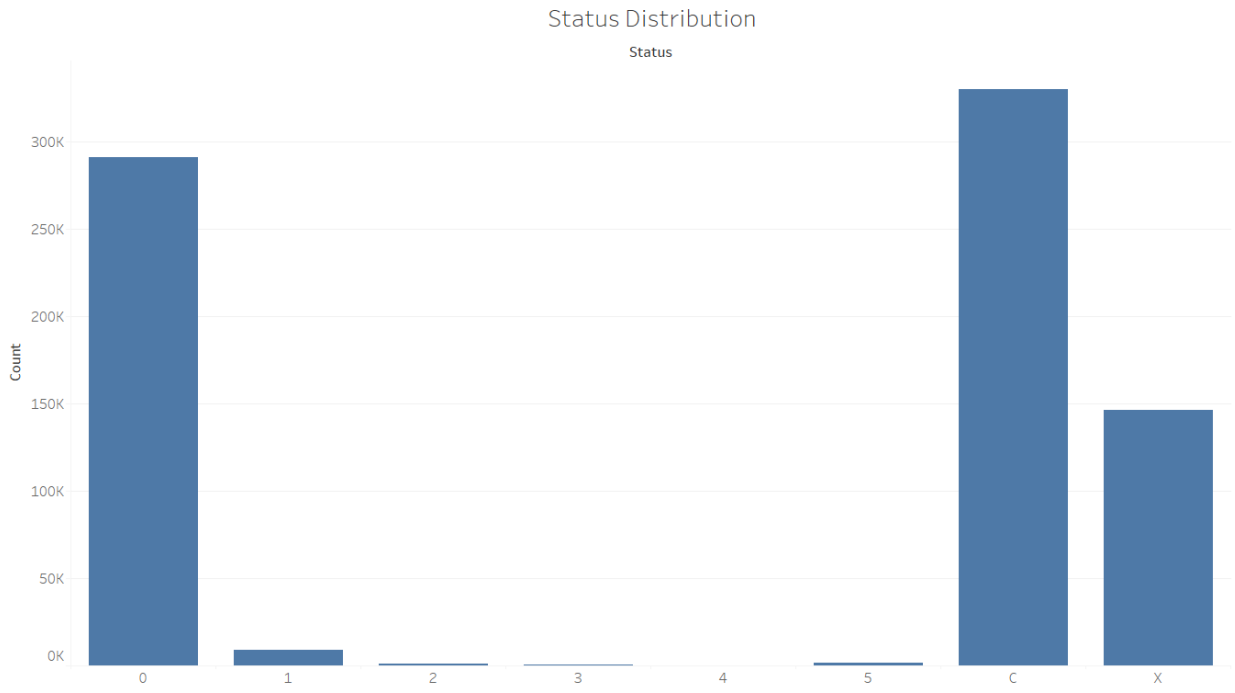
PHONE	EMAIL	FAM_SIZE	MONTHS_BALANCE	GENDER_F	GENDER_M	CAR_N	CAR_Y	REALTY_N	REALTY_Y
777715.000000	777715.000000	777715.000000	777715.000000	777715.000000	777715.000000	777715.000000	777715.000000	777715.000000	777715.000000
0.300965	0.091675	2.208837	-19.373564	0.667148	0.332852	0.608648	0.391352	0.340442	0.659558
0.458678	0.288567	0.907380	14.082208	0.471234	0.471234	0.488053	0.488053	0.473858	0.473858
0.000000	0.000000	1.000000	-60.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	2.000000	-29.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	2.000000	-17.000000	1.000000	0.000000	1.000000	0.000000	0.000000	1.000000
1.000000	0.000000	3.000000	-8.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
1.000000	1.000000	20.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Above we can see a summary of the statistical values for each numerical column. From these tables, we can see that every single person in this dataset is listed as having a mobile phone. Although it makes sense that most people would have a mobile phone considering the prevalence of these devices, it does seem odd that every single person is listed as having one. If this is incorrect, the column should be removed. If it is not incorrect, then it will not yield any meaningful results because every value is the same. Both of these possibilities point to the need to remove the phone column, so it will be removed.

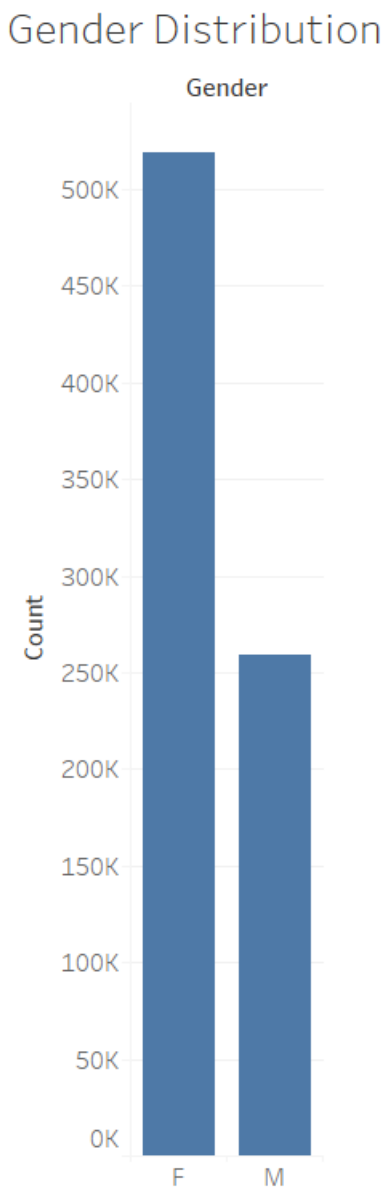
```
1 data.isnull().sum(axis = 0)
```

```
ID                        0
GENDER                    0
CAR                       0
REALTY                    0
CHILDREN                  0
INCOME                    0
INCOME_TYPE               0
EDUCATION                 0
MARITAL_STATUS            0
HOUSING_TYPE              0
DAYS_SINCE_BIRTH          0
DAYS_EMPLOYED             0
MOBILE_PHONE              0
WORK_PHONE                0
PHONE                     0
EMAIL                     0
OCCUPATION_TYPE      240048
FAM_SIZE                  0
MONTHS_BALANCE            0
STATUS                    0
dtype: int64
```

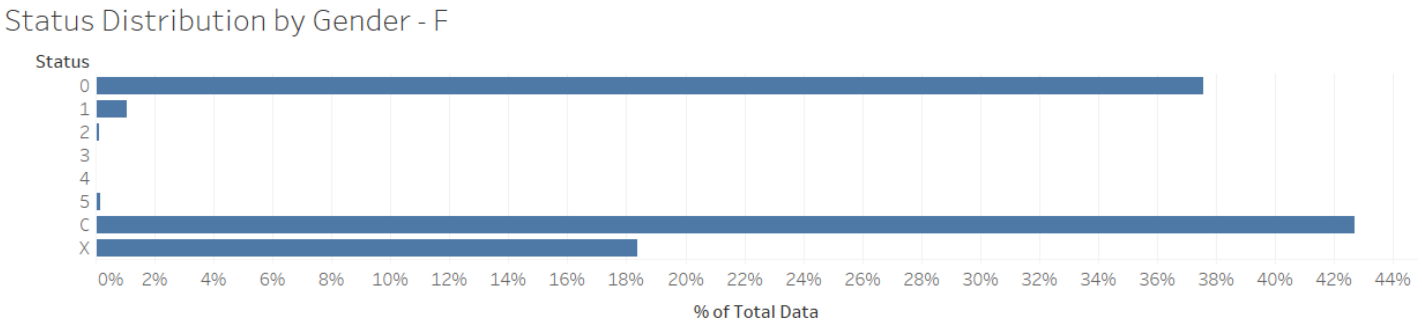
Above we can see the number of data points missing for each column. The data must have been cleaned very thoroughly, because the only missing values are in the occupation type column. One explanation for the significant number of missing values (~31% missing values) is that this column would likely be blank in the case of unemployment. This will be checked by seeing if the days employed value is positive, which would confirm that an individual is unemployed. This almost certainly will not explain all of the null values, however, because much less than thirty one percent of the population is unemployed.

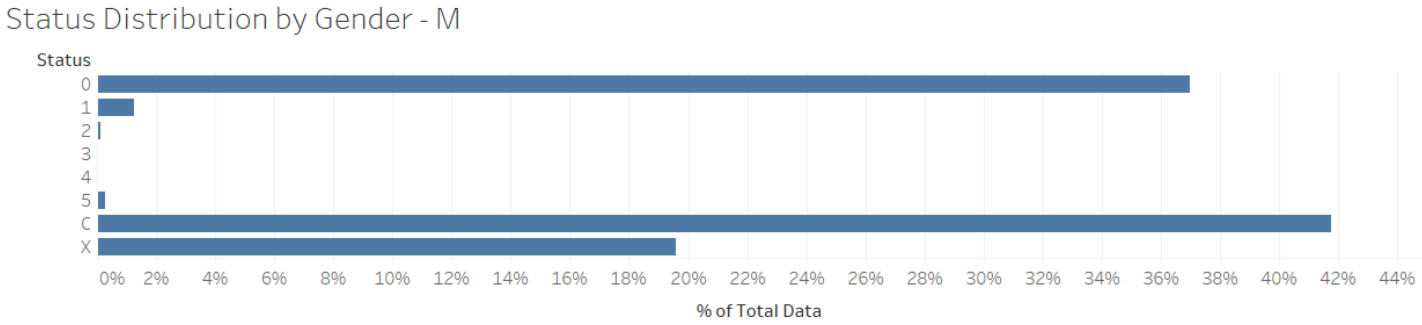


Above is a chart showing the distribution of different statuses. The most prominent status is C, which means that the debt was paid off. The second most prominent status is 0, which means that the debt is between one and twenty-nine days past due. This, although late, is often not yet considered a delinquent debt. The third most prominent status is X, which means that there was no debt that month.

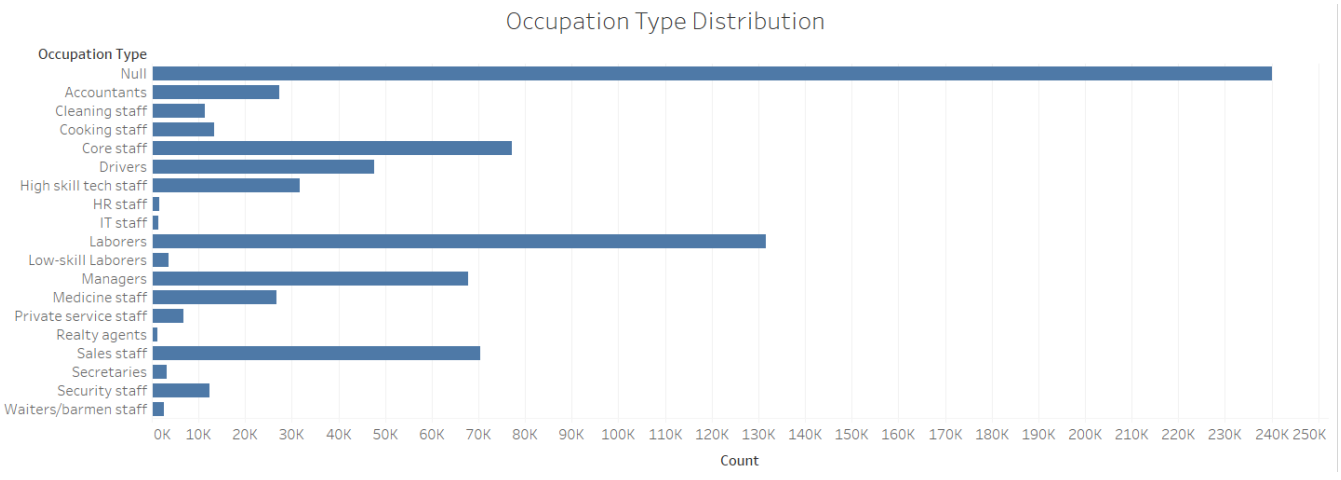


Above we see the distribution of the data between males and females. It is clear that most rows involve females, making up nearly two-thirds of the data. This should not pose any issues because there is a sufficient amount of data for both males and females.

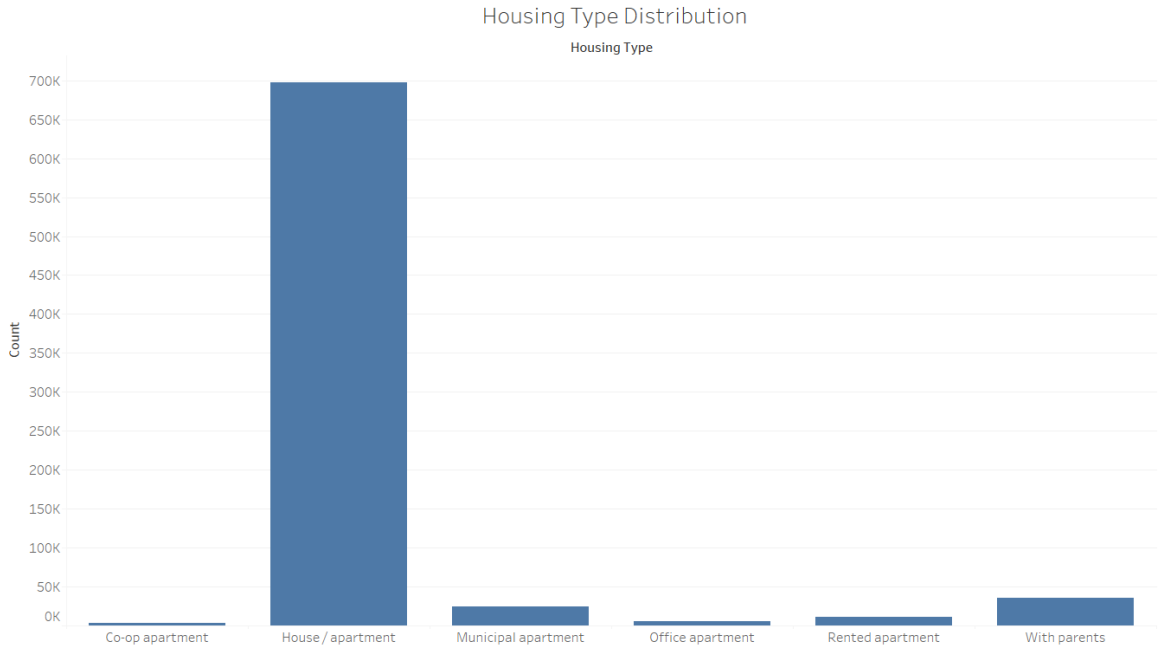




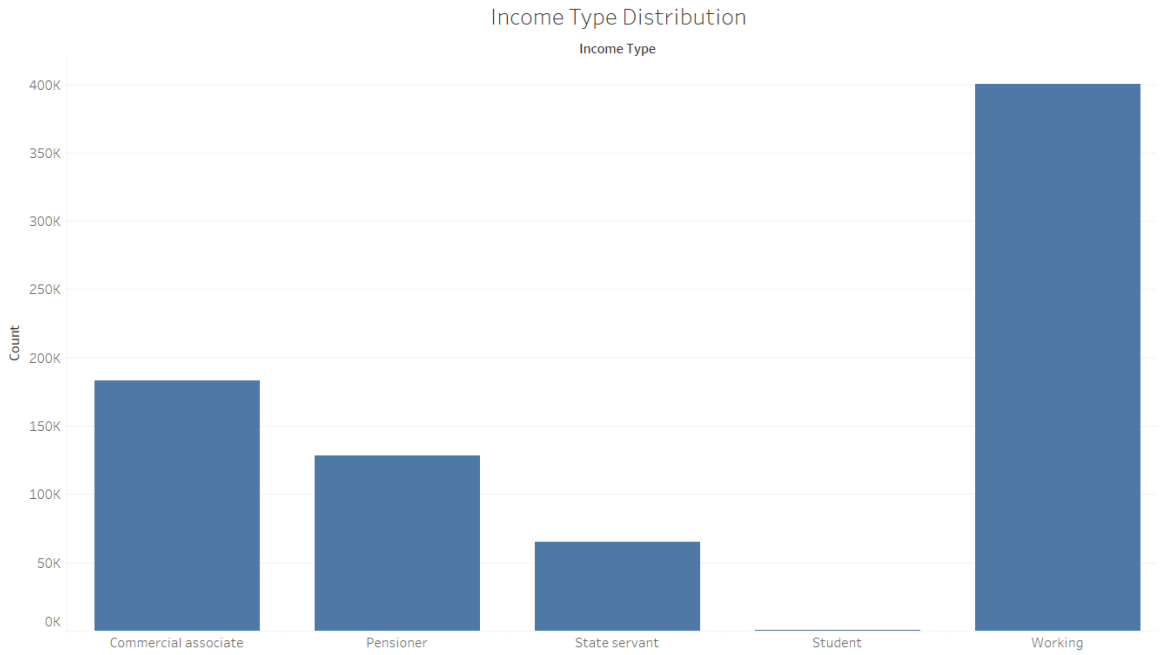
Above, we can see the distribution of data for each gender. These graphs reflect very similar payment histories for the genders, with a slightly higher rate of missed payments, and longer time since the due date, for males. This does not seem to be a large enough difference to cause concern, however.



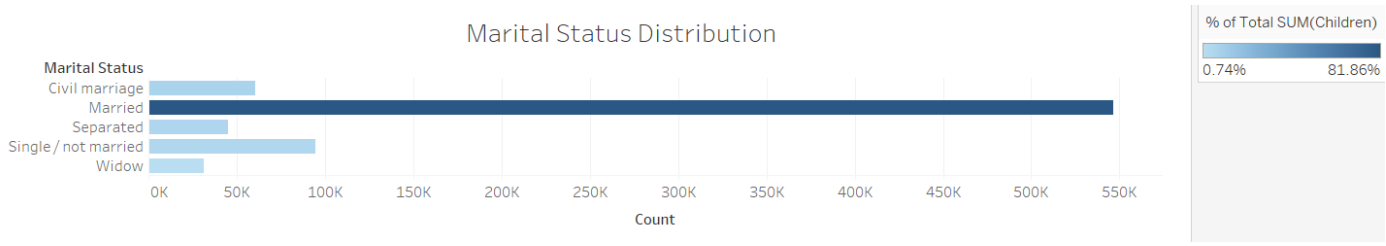
Above is a graph showing the distribution of income types. We can see that there are a disproportionate number of null values in this data (approximately 31%). This confirms what was suspected earlier: that the number of unemployed individuals might make up some portion of the null values, but there is likely another major cause.



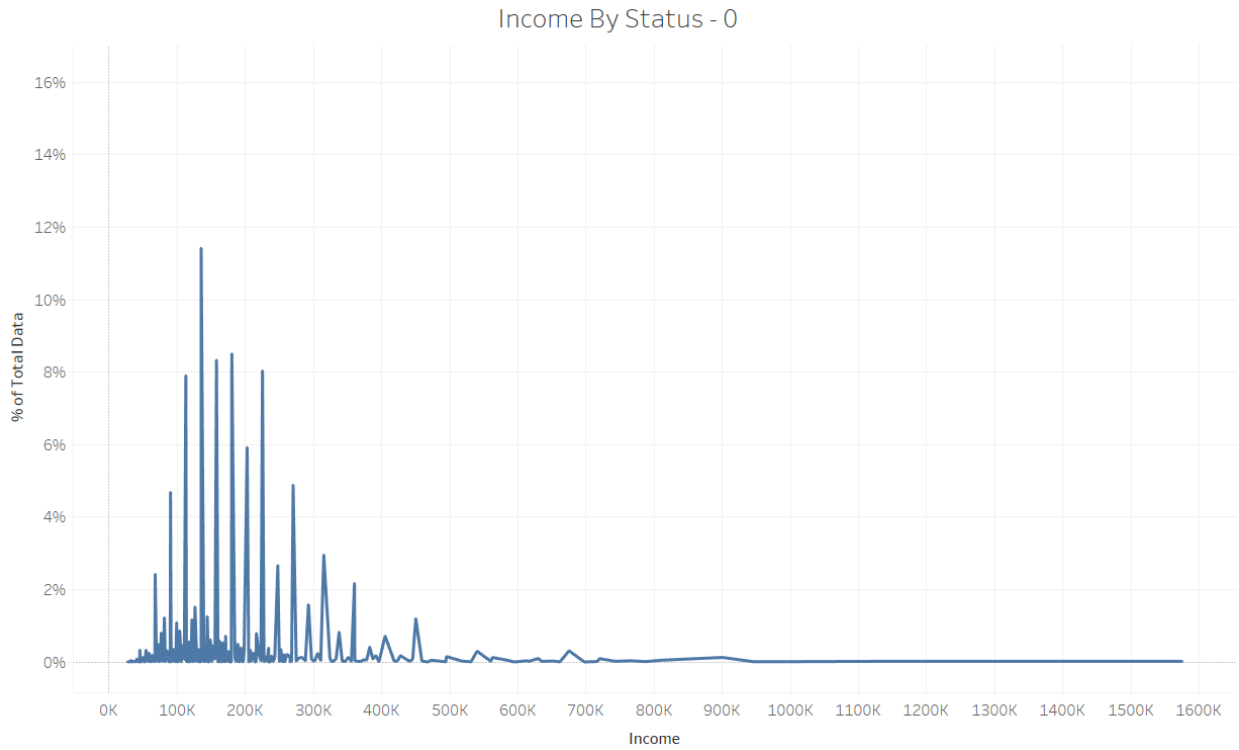
This graph shows the distribution of housing types in the data. As mentioned earlier, the most prominent output is “House / Apartment.”

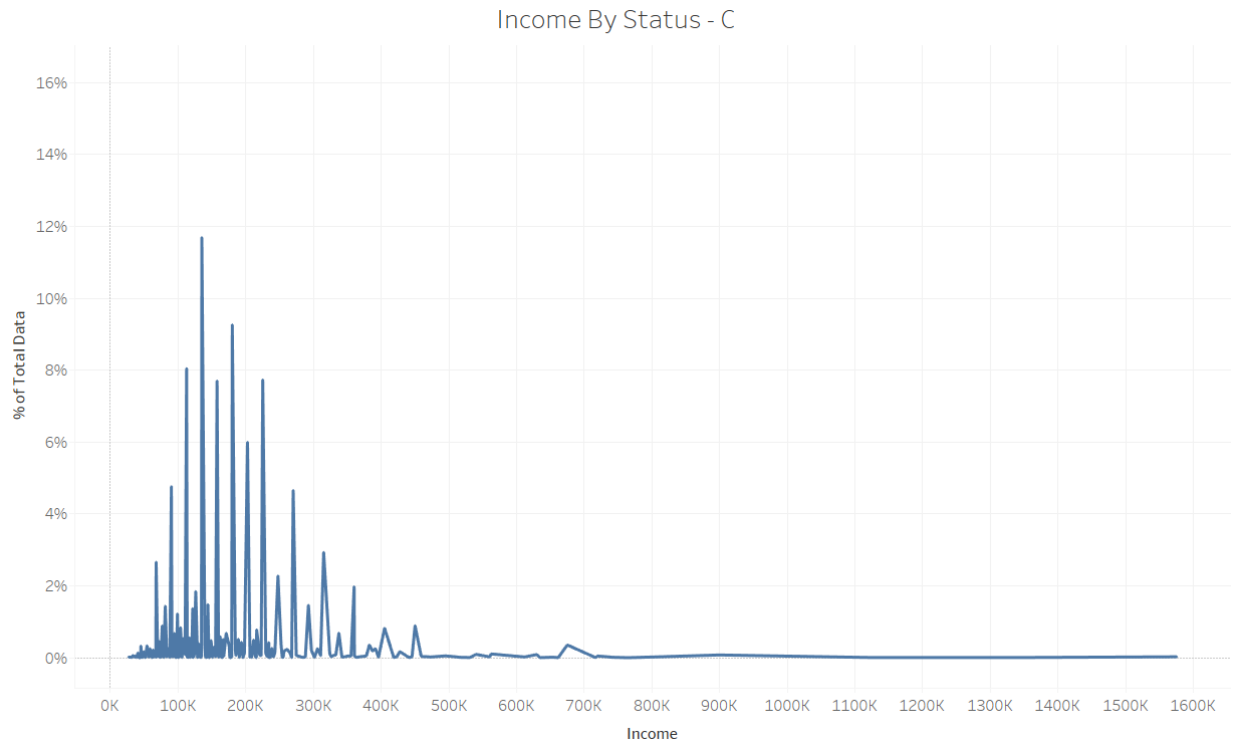


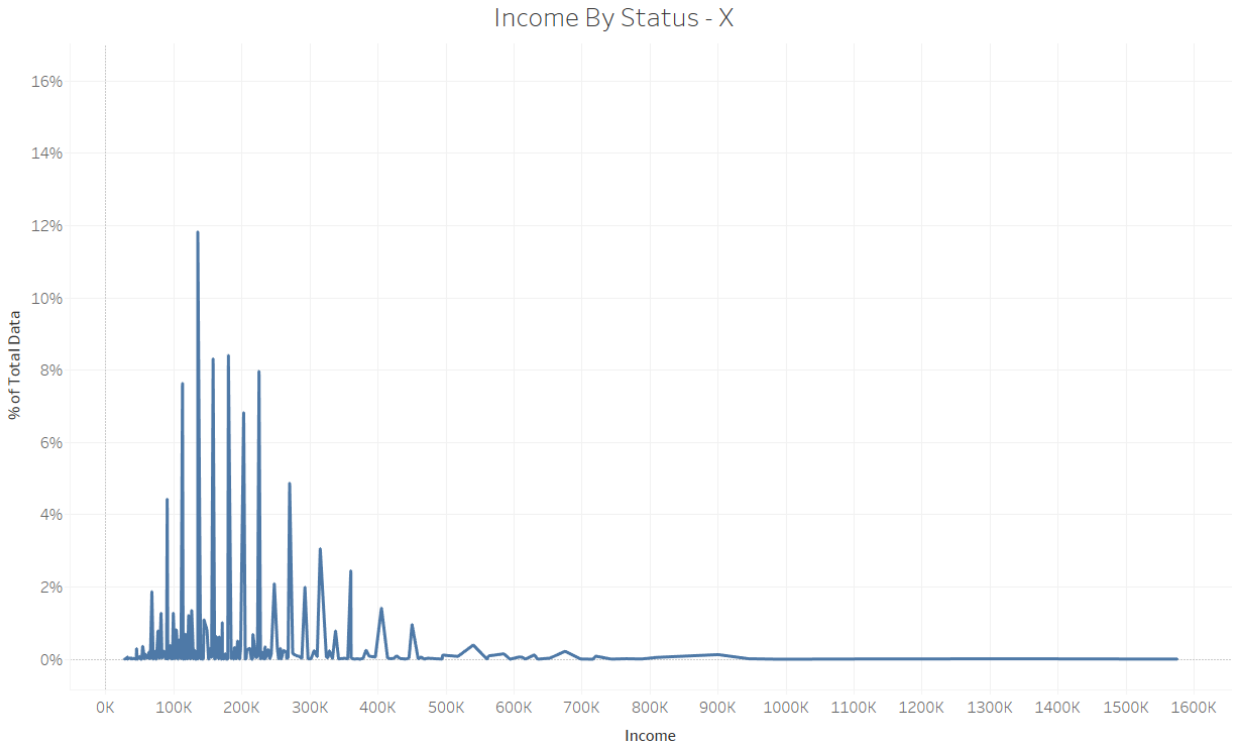
The above graph shows income type distribution. “Working”, although vague, is the most prominent income type.



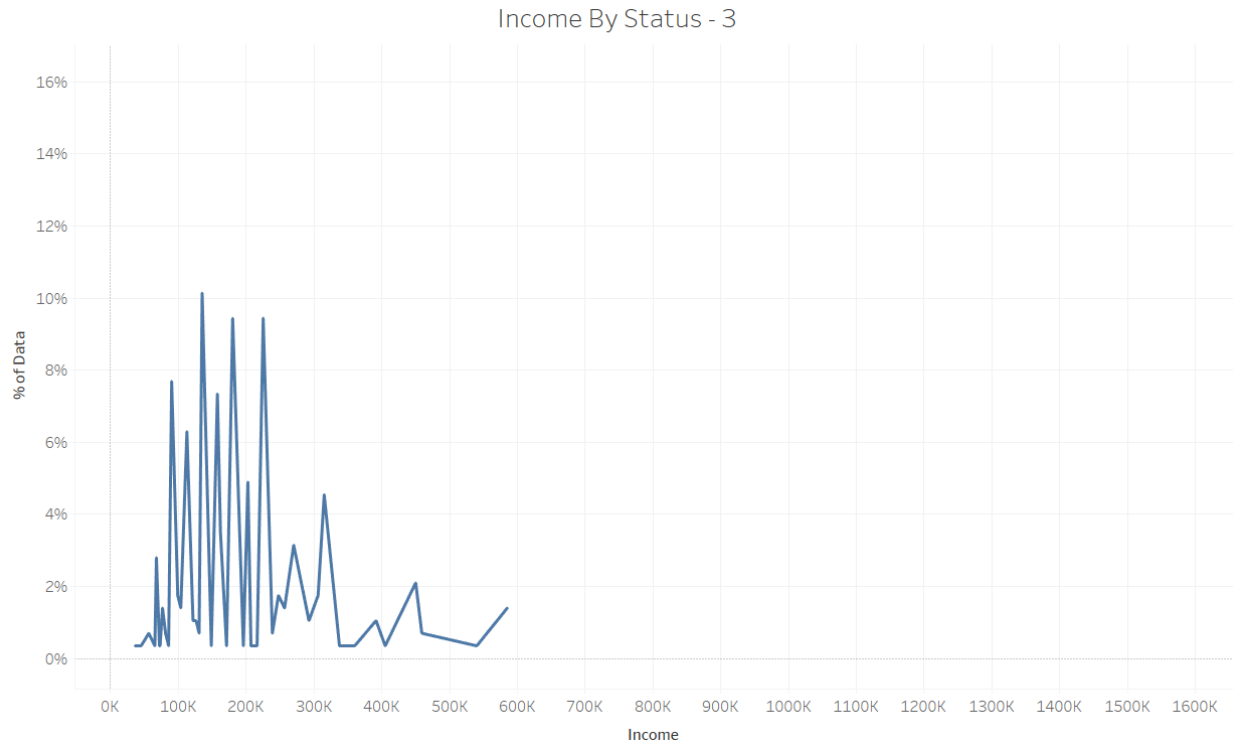
The above graph shows the distribution of the data amongst marital status. Along with this, a darker shade here signifies a higher rate of children per non-child in each group. This is not surprising, as we would reasonably expect married individuals to have more children on average than other groups.




Income By Status - C
16%
14%
12%
10%
8%
6%
4%
2%
0%
% of Total Data
0K
100K
200K
300K
400K
500K
600K
700K
800K
900K
1000K
1100K
1200K
1300K
1400K
1500K
1600K
Income


Income By Status - X
16%
14%
12%
10%
8%
6%
4%
2%
0%
% of Total Data
0K
100K
200K
300K
400K
500K
600K
700K
800K
900K
1000K
1100K
1200K
1300K
1400K
1500K
1600K
Income

Above are the line charts for the three most common status results by income. We can see that they are all relatively consistent, which makes sense considering that they can all be considered as non-late payments. The graph below shows how income is distributed in the latest recorded period of delinquency. It has some clear differences from the graphs above, although it should be noted that there is less data reflected by this graph.



In summary, the data are very clean, with very few missing values. The missing values which are present all belong to the column "Occupation Type." This can, in part, be explained by unemployment, although there seems to be a disproportionate amount of missing data for this to be the only cause. The column of "Mobile Phone" conveys that every individual has a mobile phone, which is either false or useless. In either case, this column will need to be removed. While there are over twice as many females as males in the data, and many of the categorical variables show data predominately in one of these categories, none of this seems like it will have a negative effect on results. Otherwise, the data do not show much need for adjustment before beginning the processes of machine learning.