2021

INTERMED STATS MODEL AND ANALYTICS

Group Project: Credit Card Risk Analysis- Analysis of delinquency trend for credit card repayments

Chandrika Rai
Charesh Battepati
Phu Gia Pham
Tejas Mirashi

Table of Contents

Note: There are some topics that are missed on index but are there in report.

Week 2 PPT – Chapter 1: Introduction

Business analytics problem: Define the research problem

The credit card business has been one of the major revenue sources for banks and financial institutions. Over the years, credit card has undergone significant change and it has been a widespread of practice for consumers. Due to the deregulation of the banking industry and the need of consumers, according to the *Federal Reserve Bank of Atlanta* data released in *May 2021*, 79% of consumers had at least one credit card or charge card. The constant increasing demand for credit cards, the threat of defaults on payments has been constantly increasing. The issuing banks and financial institutions are perplexed to narrow down the key indicators that represent the likelihood of default. Another major issue is the allocation of credit limit which traditionally has been linked to the level of income of the individual. Sighting the problem early, there is an urgent need for the issuers to devise a framework to detect the probability of default at the earliest. Through this project, we attempt to analyze the various factors that could result in a credit card holder defaulting on his or her credit card payments.

Why is this research problem important?

It is important to understand that the research problem is bi faceted. One of the aspects is the credit risk management for the banks. If the percentage of defaulters in this segment increases beyond control, it could have an adverse effect on the profitability. If left unaddressed, the portfolio could be damaged beyond repair and the bank / financial institution would be required to take drastic measures such as restructuring the portfolio or even halting the business for a temporary phase. The other facet is the impact on credit history of the customers. A default on credit card payments looks small in quantum, but can have a major impact on the credit history as the type of debt is unsecured in nature i.e. without any collateral. A small default now could lead the customer to be ineligible for a bigger size debt such as mortgage.

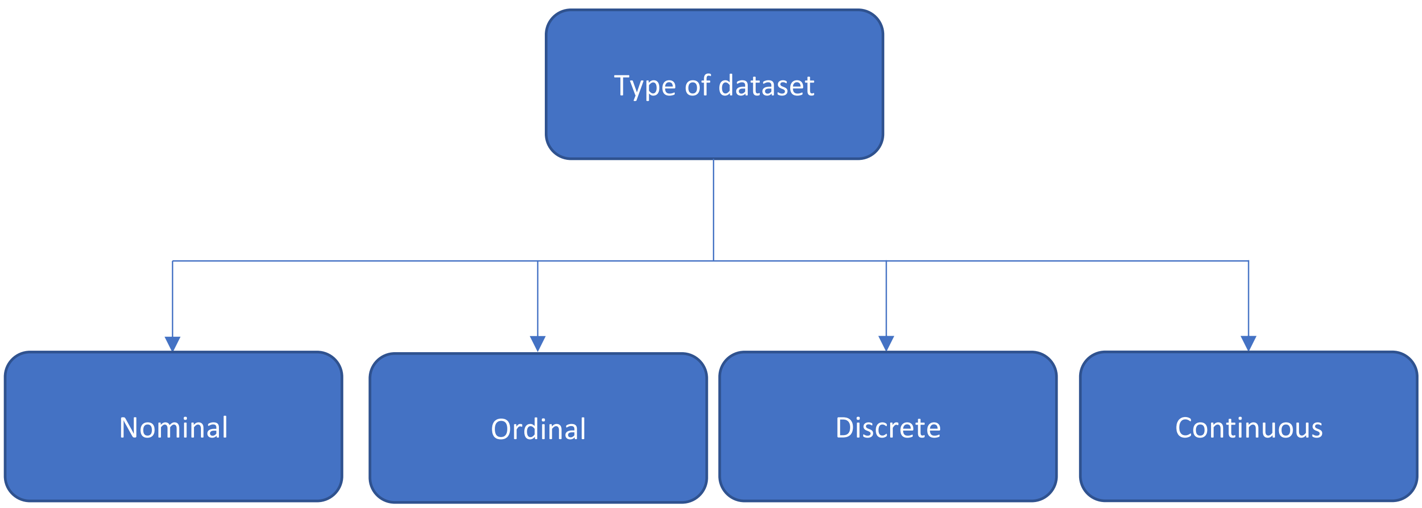
How does it relate to the STAT4600 class?

For this project, we can use knowledge that we have learned in the subject STAT4600 such as mean, median, mode, linear regression analysis, To analyze the dataset and from that, we can come up with solutions that help companies for future problems.

Methods

Describe type/source/content of data used in the project.

- Type:



Source:

The data is taken from Kaggle datasets which is a web service platform for data.

Content of data

Feature Name	**Description**	**Remarks**
ID	Client Number	Customer I'd by which the customer is represented in the financial institution.
CODE_GENDER	Gender	Gender of customer

Feature Name	Description	Remarks
FLAG*OWN*CAR	Is there a car	If customer owns a car or not
FLAG*OWN*REALTY	Is there a property	If customer owns a property or not
CNT_CHILDREN	Number of Children	The number of children the customer has
AMT*INCOME*TOTAL	Annual Income	The annual income of the customer in INR.
NAME*EDUCATION*TYPE	Education Level	The education level of customer such as academic degree, higher education, incomplete higher education, lower secondary & secondary / secondary
NAME*FAMILY*STATUS	Marital Status	Whether the customer is civil marriage, married, separated, single / not married or widow.
NAME*HOUSING*TYPE	Way of Living	The type of housing the customer resides in: co-op apartment, house / apartment, municipal apartment, office apartment, rented apartment or with parents.
DAYS_BIRTH	Age in years	Age of the customer in years.
DAYS_EMPLOYED	Duration of work in years	The number of years a customer has been employed.
FLAG_MOBIL	Is there a mobile phone	‘1’ represents customers who own a mobile phone. ‘0’ represents customers who do not own a mobile phone.
FLAG*WORK*PHONE	Is there a work phone	‘1’ represents customers who have an office phone. ‘0’ represents customers who do not have an office phone.
FLAG_PHONE	Is there a phone	‘1’ represents customers who have a landline connection.

Feature Name	Description	Remarks
		‘0’ represents customers who do not have a landline connection.
FLAG_EMAIL	Is there an email	‘1’ represents customers who have a working e-mail id. ‘0’ represents customers who do not have a working e-mail id.
JOB	Job	The occupation of the customer.
BEGIN_MONTHS	Record month	The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on
STATUS	Status	0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days overdue 3: 90-119 days overdue 4: 120-149 days overdue 5: Overdue or bad debts, write-offs for more than 150 days C: paid off that month X: No loan for the month
TARGET	Target	Risk users are marked as '1', else are '0'

Describe the analysis and modeling methods

To analyze the data, we will utilize descriptive analysis, diagnostic analysis, and prescriptive analysis as we are going to determine the mean and median of each credit card holder's annual income, their ownership, and more independent variables. Charts, graphs, and tables will be created to visualize data on the dependent variables and independent variables by grouping the data into classes.

Next, utilizing the data visualization, we will provide suggestions for banking systems. For analysis, R and Python programming languages will be utilized for this project.
Analyze various models as follows:

- Linear regression model
- Multi Linear regression
- Exploratory data analysis

The software used:

In this research, the programming language used is Python. Python is a high-level programming language, invented by Guido Van Rossum. This language is very popular because of its code readability and compact line of code.

In addition, Python is an open source and a syntactically simple programming language with rich, thriving community supported from lots of people and experts in all over the world. Python is also used for various applications such as natural language processing, machine learning, data analytics.

R Studio and Excel will be used for handling the dataset.

Methods

Chapter 1:

Example of Element or member, variable and observation in the dataset

An element or member of credit card delinquency analysis is a specific subject or object about which the information is collected.

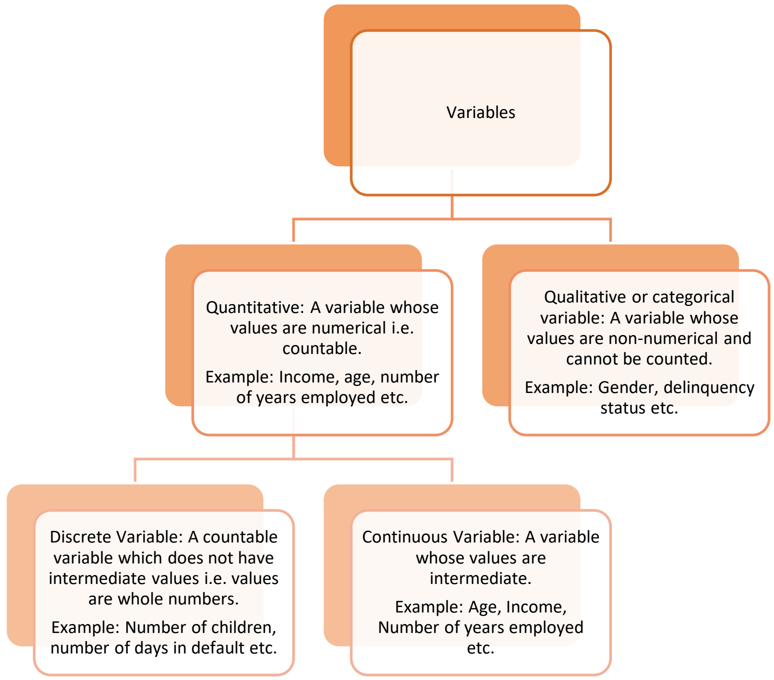
Variables

Gender	**Number of children**	**Income**	**Age**
Female	2+ children	270,000	36.32
Female	No children	81,000	48.98
Male	No children	270,000	53.64
Female	1 child	112,500	41.39

An element or member

An observation or measurement

Types of variables – Discrete and continuous



Discuss Cross-section and time series in dataset

Cross-section data: Cross-section data refers to data collected on different elements at the same point or same period. In our dataset, the author has collected data for delinquency on a particular date (the date is not mentioned as it may lead to breach of privacy between the bank and the customers).

Gender	Income	Begin months	Status
Female	270,000	6	C
Female	81,000	4	0
Male	270,000	0	C
Female	112,500	3	0

Time-series data: Time-series data refers to data collected on the same element for the same variable at different points of time. As mentioned earlier, the author of our dataset has recorded the delinquency status at the end of a particular month and hence, our dataset does not have time-series data.

Discuss Population and sample in dataset

Population: Population refers to the study of all the variables, elements and observations present in the dataset. A study of all the observations pertaining to all the customers can be referred to as analysis of population.

Sample: The study of variables pertaining to delinquent customers alone can be referred as analysis of sample. In simple terms, sample can be referred to as a subset of the dataset.

Discuss Census and survey sample in dataset

Census: Census can be referred to as a survey of all members of a population. In the dataset, survey on all the customers of the bank refers to census.

Sample survey: Sample survey can be referred to as survey of a portion of the population. In the dataset, survey of all delinquent customers can be referred to as sample survey.

Pick Representative sample in dataset

Representative sample: A sample picked up from the delinquent customers can be referred to as representative sample. The traits of this sample can be analyzed and compared to the traits of other delinquent customers from the population. For example, Customer ID 5022428 is a delinquent customer and has been categorized as risky by the bank. The traits of this customer can be analyzed with other delinquent customers to derive insights on common traits.

Construct Two types of sampling in dataset

Sampling with replacement: Selecting a customer from the population and then putting it back in the population for the next selection is known as sampling with replacement. Since, the population contains the same number of observations after every selection, the same observation can be selected multiple times.

Customer ID	Gender	Income	Begin months	Status
5065438	Female	270,000	6	C
5142753	Female	81,000	4	0
5111146	Male	270,000	0	C
5010310	Female	112,500	3	0

In the above example, let us say that Customer ID 5065438 is selected for analysis. In sampling with replacement, the observation is put back in the population for next selection and has a possibility of being selected again.

Sampling without replacement: If a selection of a customer is made and is not put back in the population, the population size reduces with each selection. Hence, there is no possibility that a same observation will be selected twice.

Customer ID	Gender	Income	Begin months	Status
5065438	Female	270,000	6	C
5142753	Female	81,000	4	0
5111146	Male	270,000	0	C
5010310	Female	112,500	3	0

In the above example, let us say that Customer ID 5065438 is selected for analysis. In sampling without replacement, the observation is not put back in the population for next selection and resultantly cannot be selected again.

Construct Random sample and Non-random sample in dataset

Random sample: Random sample refers to selecting any observation without applying any pre-determined filters. Hence, every observation has an equal probability of selection. For example,

Customer ID	Gender	Income	Begin months	Status	Target
5065438	Female	270,000	6	C	0
5142753	Female	81,000	4	0	0
5111146	Male	270,000	0	C	0
5010310	Female	112,500	3	0	0

In the above sampling, any four random customers are selected.

Non-random sample: Non-random sample refers to selecting an observation as per certain pre-determined filters. Hence, every observation does not have an equal probability of selection. For example,

Customer ID	Gender	Income	Begin months	Status	Target
5022428	Female	90,000	38	2	1
5092251	Male	112,500	44	2	1
5029719	Female	265,500	9	2	1
5115608	Female	202,500	16	5	1

In the above sampling, only those customers who are categorized as risky are being for sampling. Hence, non-delinquent customers have no probability of selection.

Discuss Sampling error and non-sampling error in dataset

All the customers who are delinquent own a mobile phone. Hence, from the population, this is known as **sampling error.**

Data entered for a particular customer by the bank is incorrect. Such kind of error is known as **non-sampling error.**

Types of errors

Selection error: A sample is constructed from all delinquent customers to check those owning a mobile phone. On analysis, we find that all the delinquent customers own a mobile phone. Analysis leads us to the conclusion that this variable is to be dropped for regression and selection of observation keeping mobile phone as pivot can lead to a selection error.

Non-response error: Non-response error is where a customer does not respond to a question asked in a survey. Since, our dataset is received from the bank directly, we do not have any non-response error in sampling.

Response error: A response error is where a customer does not respond to a question asked in a survey correctly. Since, our dataset is received from the bank directly, we do not have any response error in sampling.

Voluntary response error: A voluntary response error is where a customer chooses to respond to a question or not. Since, our dataset is received from the bank directly, we do not have any voluntary response error in sampling.

Create a sample using simple random sampling

	ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL
338603	5042244	M	N	Y	No children	90000.0
13238	5047849	M	Y	Y	No children	315000.0
47429	5096139	M	N	N	No children	135000.0
515245	5105127	M	Y	Y	No children	1575000.0
529532	5037047	M	Y	Y	No children	135000.0

NAME_EDUCATION TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	AGES
Secondary/secondary special	Civil marriage	House/apartment	41.48
Secondary/secondary special	Married	House/apartment	59.92
Secondary/secondary special	Married	House/apartment	50.72
Secondary/secondary special	Married	House/apartment	29.17
Secondary/secondary special	Married	With parents	44.58

FLAG_MOBIL	FLAG_WORK_PHONE	FLAG_PHONE	FLAG_EMAIL
1	1	1	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0

JOB	BEGIN_MONTHS	STATUS	TARGET
Low-skill Laborers	10	0	0
Drivers	13	X	0
Drivers	29	X	0
Laborers	13	C	0
Core staff	5	C	0

(See Appendix 1)

Create a sample using systematic random sampling

Systematic sampling for selecting every 5th row

	ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN
0	5065438	F	Y	N	2+ children
5	5067057	F	Y	Y	No children
10	5026032	M	Y	Y	No children
15	5023566	M	Y	Y	2+ children
20	5022044	F	Y	N	1 children

AMT_INCOME_TOTAL	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	AGES
270000.0	Secondary/secondary special	Married	With parents	36.32
144000.0	Secondary/secondary special	Married	House/apartment	42.18
99000.0	Secondary/secondary special	Married	Rented apartment	27.20
270000.0	Secondary/secondary special	Married	House/apartment	33.76
225000.0	Secondary/secondary special	Separated	House/apartment	32.09

NUMBER_OF_YEARS_EMPLOYED	FLAG_MOBIL	FLAG_WORK_PHONE	FLAG_PHONE	FLAG_EMAIL
6.30	1	0	0	0
8.11	1	0	1	0
4.19	1	0	0	0
3.06	1	0	1	0

4.77	1	0	0	1

JOB	BEGIN_MONTHS	STATUS	TARGET
Managers	6	C	0
Core staff	25	0	0
Managers	15	0	0
Laborers	12	C	0
Core staff	25	C	0

See Appendix 2

Create a sample using stratified random sampling

Defined a dictionary for creating a strata of female customers, with no children, working as manager and non-risky customers and tagged then with the ID.

	ID	CODE_GENDER	CNT_CHILDREN	JOB	TARGET
0	5036943	F	No children	Manager	0
1	5023870	F	No children	Manager	0
2	5021648	F	No children	Manager	0
3	5112626	F	No children	Manager	0
4	5089398	F	No children	Manager	0

See Appendix 3

a. Create a sample using cluster sampling

	ID
0	5000000
1	5000001
2	5000002
3	5000003
4	5000004
...	...
999994	5999994
999995	5999995
999996	5999996
999997	5999997
999998	5999998

999999 rows × 1 columns

Now that the IDs are mapped in order, we can further map the customers according to ID.

See Appendix 4

Discuss if dataset is observational study or controlled experiment

The dataset is observational study as we are merely using the data points of bank customers to predict probable delinquent customers without putting any variable or observation to any kind of conditional experiment.

The dataset currently contains a delinquency trend which can be used to predict customers which can probably turn delinquent in future.

Chapter 2

Create frequency distribution table for qualitative variables

Job	**Frequency Distribution**
Accountants	27223
Cleaning staff	11399
Cooking staff	13416
Core staff	77112
Drivers	47678
HR staff	1686
High skill tech staff	31768
IT staff	1319
Laborers	131572
Low-skill Laborers	3623
Managers	67738
Medicine staff	26691
Private service staff	6714
Realty agents	1260
Sales staff	70362
Secretaries	3149
Security staff	12400
Waiters/barmen staff	2557

Here, we have made a table to display the count of customers in each job. One thing to be observed is that data is arranged in alphabetical order. Count wise, most customers are laborers, and the least are realty agents.

See Appendix 5

Determine the relative frequency and percentage of above-mentioned qualitative variables

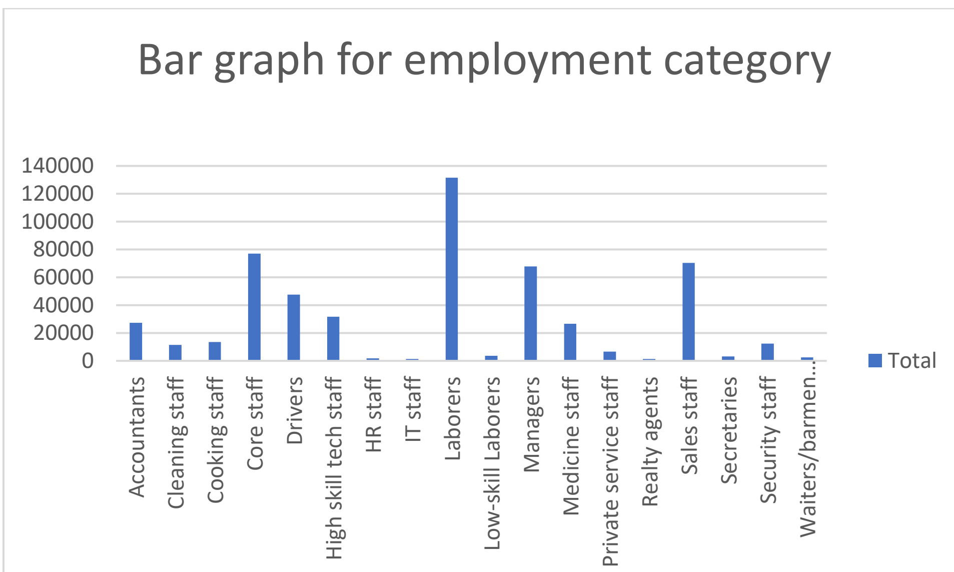
Laborers	0.244709
Core staff	0.143420
Sales staff	0.130865

Managers	0.125985
Drivers	0.088676
High skill tech staff	0.059085
Accountants	0.050632
Medicine staff	0.049642
Cooking staff	0.024952
Security staff	0.023063
Cleaning staff	0.021201
Private service staff	0.012487
Low-skill Laborers	0.006738
Secretaries	0.005857
Waiters/barmen staff	0.004756
HR staff	0.003136
IT staff	0.002453
Realty agents	0.002343

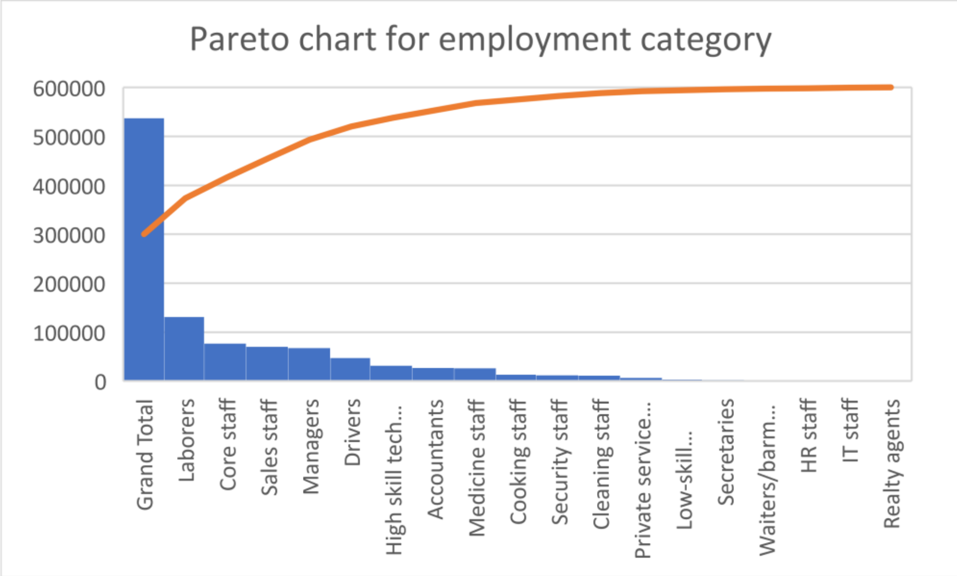
See Appendix 6

Create bar graphs and Pareto charts for above mentioned variables

Bar graph for number of customers employed in each category



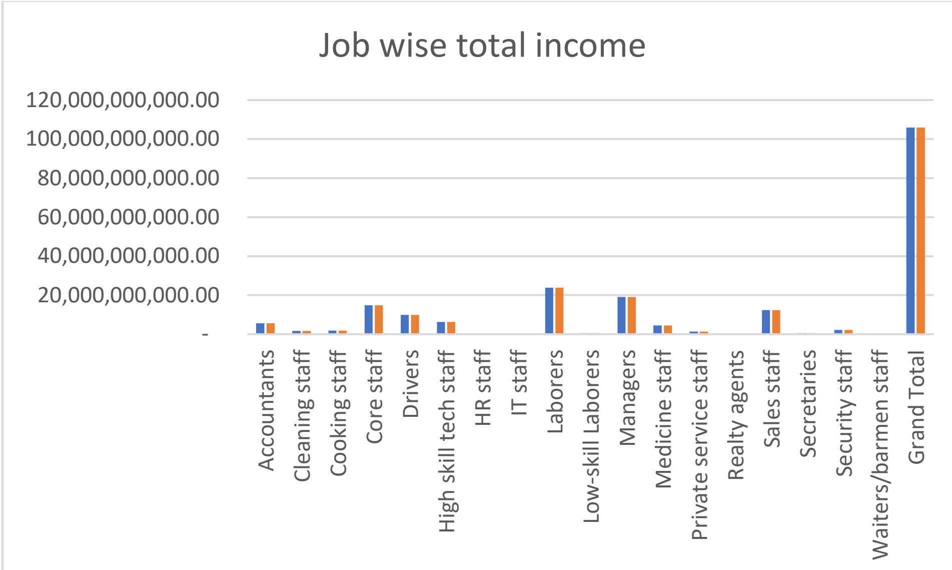
Pareto chart for number of customers employed in each category



a. Create frequency distribution table for quantitative variables in dataset (less than method on page 32 and single-value classes on page 34)
   Since our dataset is large to use less-than or single values, we have created frequency distribution table using job wise income.

Job	Total Income
Accountants	5,655,891,577.50
Cleaning staff	1,699,544,700.00
Cooking staff	1,934,445,150.00
Core staff	14,799,151,737.00
Drivers	9,963,220,500.00
High skill tech staff	6,288,599,250.00
HR staff	331,645,500.00
IT staff	294,920,100.00
Laborers	23,907,711,150.00
Low-skill Laborers	462,091,500.00
Managers	19,082,742,660.00
Medicine staff	4,466,669,121.00
Private service staff	1,379,547,000.00
Realty agents	306,274,500.00
Sales staff	12,304,794,703.50
Secretaries	513,936,000.00
Security staff	2,199,033,000.00
Waiters/barmen staff	393,156,000.00
Grand Total	**105,983,374,149.00**

Create histogram for quantitative variables in dataset



Create the cumulative frequency distribution table with cumulative relative frequency and cumulative percentage for the quantitative variables in your dataset.

Job	Total Income	Relative frequency	Percentage
Accountants	5655891578	5655891578	0.053
Cleaning staff	1699544700	7355436278	0.068
Cooking staff	1934445150	9054980978	0.084
Core staff	14799151737	10989426128	0.102
Drivers	9963220500	25788577865	0.239
High skill tech staff	6288599250	35751798365	0.332
HR staff	331645500	42040397615	0.390
IT staff	294920100	42372043115	0.393
Laborers	23907711150	42666963215	0.396
Low-skill Laborers	462091500	66574674365	0.618
Managers	19082742660	67036765865	0.623
Medicine staff	4466669121	86119508525	0.800

Private service staff	1379547000	90586177646	0.841
Realty agents	306274500	91965724646	0.854
Sales staff	12304794704	92271999146	0.857
Secretaries	513936000	104,576,793,849.00	0.971
Security staff	2199033000	105,090,729,849.00	0.976
Waiters/barmen staff	393156000	107,289,762,849.00	0.996
Grand Total	105,983,374,149.00	107,682,918,849.00	1.000

Discuss the shapes of histogram

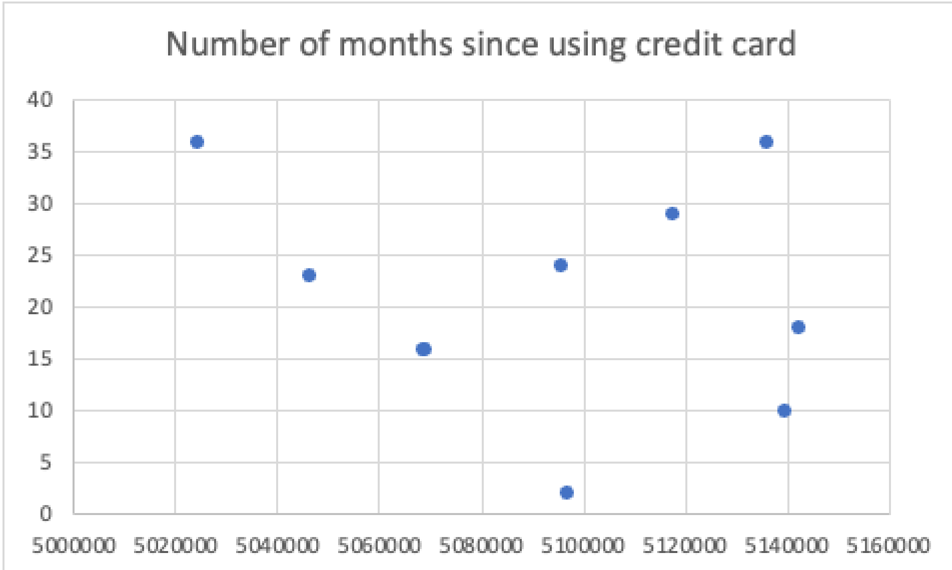
Both the created histograms are almost symmetric. The reason for this is maximum number of customers are laborers and are in the middle of the histogram.

Randomly select 10 observations from your dataset to create a stem-and-leaf display

ID	BEGIN_MONTHS
5096800	2
5139478	10
5068679	16
5068432	16
5142196	18
5046178	23
5095399	24
5117593	29
5135783	36
5023963	36

Stem	Leaf
0	2
1	0 6 6 8
2	3 4 9
3	6 6

Randomly select 10 observations from your dataset to create dot plots



Chapter 3

Calculate the mean of each variable in your dataset

Variables	Mean
AMT_INCOME_TOTAL	197117.13
AGE	41.13
NUMBER_OF_YEARS_EMPLOYED	7.57
FLAG_MOBIL	1
FLAG_WORK_PHONE	0.28
FLAG_EMAIL	0.1
BEGIN_MONTHS	19.31
TARGET	0

See Appendix 7

If your dataset has outliers, please try the k% trimmed mean

There is notably a large difference between the 75th %tile and max values of the variables Annual Income, Number of Years Employed, and Begin months. This means that we have a lot of outliers in these 3 variables.

For these 3 variables, we will try the trimmed mean at 10%:

Variables	Trimmed mean
AMT_INCOME_TOTAL	183963.73
NUMBER_OF_YEARS_EMPLOYED	6.522
BEGIN_MONTHS	18.1501

See Appendix 8

Calculate the median of each variable in your dataset

Variables	Median
AMT_INCOME_TOTAL	180000
AGE	40.51
NUMBER_OF_DAYS_EMPLOYED	5.88
FLAG_MOBIL	1
FLAG_WORK_PHONE	0
FLAG_EMAIL	0
BEGIN_MONTHS	17
TARGET	0

See Appendix 7

Discuss the different between mean and median

The mean is the average in which the sums of all the elements is divided by the total numbers of the elements.

The median is the middle value in the list of given numbers in increasing order.

Discuss the mode of each variable in dataset

The mode is the value that occurs with the highest frequency in the dataset

Variables	Mode	Discussion
CODE_GENDER	F	It means that in the dataset, there are more females than males
FLAG_OWN_CAR	N	It means that the number of people who do not own cars is larger than the number of people who own car
FLAG_OWN_REALTY	Y	It means that the number of people who own properties is larger than the number of people who do not own property
CNT_CHILDREN	No children	It means that most people in the dataset do not have any children
NAME_EDUCATION_TYPE	Secondary/Secondary special	It means that most people education level

		in the dataset is Secondary/Secondary special (low education level)
NAME_FAMILY_STATUS	Married	It means that the number of married people in the dataset is the largest
NAME_HOUSING_TYPE	House/apartment	It means that most people in the dataset live in House/apartment
AGES	34.73	It means that most people are around 34 years old
NUMBER_OF_YEARS_EMPLOYED	1.18	It means that most people just started working for around a year
FLAG_MOBIL	1	It means most people in the dataset have mobile phone
FLAG_WORK_PHONE	0	It means that most people in the dataset do not have work phone
FLAG_PHONE	0	It means that most people in the dataset do not have work phone
FLAG_EMAIL	0	It means that most people in the dataset do not have email
JOB	Laborers	The percentage of people who work as laborers is highest
BEGIN_MONTHS	1	It means most of them are new with credit card
STATUS	C	It means most people in the dataset have clear their balance in their credit card
TARGET	0	It means most people are in non-default status

See Appendix 9

Discuss if each variable is unimodal/bimodal/multimodal

According to the result above, each variable only has one mode. Therefore, all the variables are unimodal.

Use 10 observations from dataset to create weighted mean

In this case, we use the first 10 observations to calculate the weighted mean of the two variables Annual Income and Ages.

The weighted mean = $\frac{\Sigma wx}{\Sigma x}$ = 203677.54

See Appendix 10

Discuss the relationship among the mean, median, and mode of two variables that you randomly pick from your dataset

	Annual Income	Number of Years Employed
Mean	197117.13	7.57
Median	180000.00	5.88
Mode	135000.00	1.18

As can be seen, the mean is the largest, the mode is the smallest in both variables. Therefore, the histogram of these 2 variables will be skewed to the right or be positive skewed. For these two situations, the mean is the largest because it is affected by outliers and it is obviously that in both variables, there are a lot of outliers.

See Appendix 7 and 9

Calculate the range for each variable in your dataset

Range = Largest value – Smallest value

	Largest Value	Smallest Value	Range
AMT_INCOME_TOTAL	1575000	27000	1548000
AGES	67.43	20.52	46.91
NUMBER_OF_DAYS_EMPLOYED	43.05	0.05	43
FLAG_MOBIL	1	1	0
FLAG_WORK_PHONE	1	0	1
FLAG_PHONE	1	0	1
FLAG_EMAIL	1	0	1
BEGIN_MONTHS	60.00	0	0
TARGET	1	0	1

Calculate the variance and standard deviation for each variable in your dataset and interpret the results

	Standard deviation	Variance

AMT_INCOME_TOTAL	104138.96	10844922989.88
AGES	9.36	87.61
NUMBER_OF_DAYS_EMPLOYED	6.56	43.03
FLAG_MOBIL	0	0
FLAG_WORK_PHONE	0.45	0.2025
FLAG_PHONE	0.46	0.2116
FLAG_EMAIL	0.3	0.09
BEGIN_MONTHS	14.04	197.12
TARGET	0.06	0.0036

Compare the standard deviation with the mean, for the annual income and the begin month of working, these two variables have outliers. Therefore, the standard deviation of these two variables will be increased. Then the values of the dataset are spread over a relatively larger range around the mean.

See Appendix 7

a. Calculate the coefficient of variation for two different variables in your dataset
Coefficient of Variation (CV) = $\frac{\sigma}{\mu}$ * 100%

	Standard deviation	Mean	Coefficient of Variation
AMT_INCOME_TOTAL	104138.96	197117.13	52.83%
AGES	9.36	41.13	22.76%
NUMBER_OF_DAYS_EMPLOYED	6.56	7.57	86.66%
FLAG_MOBIL	0	1	0%
FLAG_WORK_PHONE	0.45	0.28	160.71%
FLAG_PHONE	0.46	0.30	153.33%
FLAG_EMAIL	0.3	0.1	300%
BEGIN_MONTHS	14.04	19.31	72.71
TARGET	0.06	0.00	infinity

Create an example similar to Example 3-21 using your dataset and use Chebyshev's theorem to solve the problem

The average age of people in the dataset was found to be 41.13 with a standard deviation of 9.36. Using Chebyshev's theorem, find the minimum percentage of people in this dataset who are between the age of 21.13 and 61.13

Mean = 41.13

21.13 <----------------------> 41.13 <----------------------> 61.13
20 20

Standard deviation = 9.36

K = 20/9.36 = 2.14 ~ 2

Therefore, at least 75% of the people in this dataset who are between the age of 21.13 and 61.13

Discuss the empirical rule of your dataset using one variable. You can refer to the example 3-22 on page 50

In our dataset, there is no variable which can apply empirical rules. According to empirical rule, the mean, mode, and median should be equal but, in our dataset, there is no such things.

Please calculate the percentile of 20 observations using one variable

We are using the first 20 observations and Annual Income variable to calculate the 25th, 50th, 75th percentile

25th percentile = 129375
50th percentile = 157500
75th percentile = 236250

See Appendix 11

Calculate the percentile rank of 20 observations using one variable

We using the first 20 observations of Annual Income to calculate percentile rank

	Annual Income	Percentile Rank
1	270000.0	0.826551
2	81000.0	0.044168
3	270000.0	0.826551
4	112500.0	0.153697
5	139500.0	0.348847
6	144000.0	0.357172
7	180000.0	0.524616
8	405000.0	0.964870
9	135000.0	0.288680
10	270000.0	0.826551
11	99000.0	0.099121
12	103500.0	0.107639
13	225000.0	0.707055
14	171000.0	0.469066
15	135000.0	0.288680
16	270000.0	0.826551
17	225000.0	0.707055
18	202500.0	0.617477
19	135000.0	0.288680
20	67500.0	0.019125

See Appendix 12

Calculate the quartiles and IQR of 20 observations using one variable

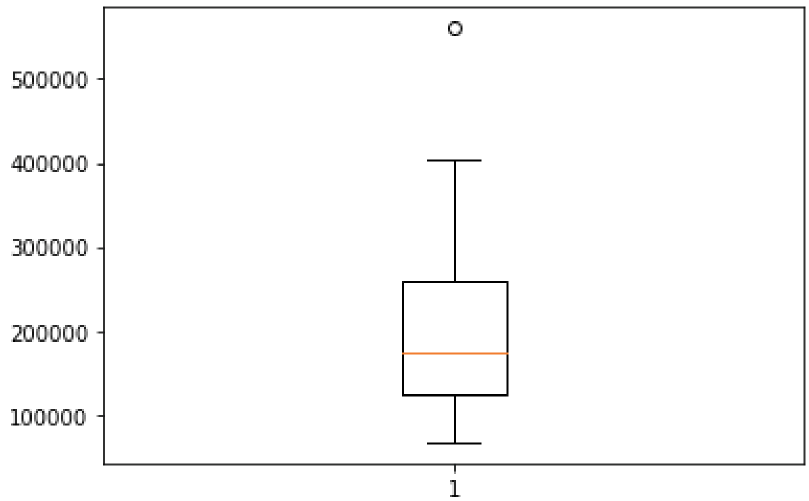
We are using the first 20 observations and calculate the quartiles for Annual Income
According to the definition:
1st quartile = 25th percentile = 129375

2^{nd} quartile = 50^{th} percentile = 157500
3^{rd} quartile = 75^{th} percentile = 236250

Therefore, IQR = Q3 – Q1 = 236250 – 129375 = 106875

Create the box and whisker plot of 30 observations using one variable

We are using the first 30 observation of Annual Income variable to create the box and whisker plot



See Appendix 13

WEEK - 6 (Chapter 4: Probability)

Experiment, Outcome, and Sample Space :-

An **experiment** is a process that, when performed, results in one and only one of many observations.

These observations are called the **outcomes** of the experiment.

The collection of all outcomes for an experiment is called a **sample space.**

1. One example for experiment, outcome, and sample space.

- **Experiment** : Ask for the marital status the participant
- **Outcomes** : married or not married
- **Sample Space** : married and unmarried.

Simple Event :-

An event is a collection of one or more of the outcomes of an experiment.

An event that includes one and only one of the (final) outcomes for an experiment is called a simple event and is usually denoted by *Ei* .

2. Give out one example for event and simple event.

- Example for an event is age.
- Example for simple event is house owned or rented.

Compound Event :-

A compound event is a collection of more than one outcome for an experiment.

3. Give one example for the compound event.

- The example for compound event is profession.

Mutually Exclusive Events, Independent vs Dependent Events :-

Events that cannot occur together are said to be mutually exclusive events.

Two events are said to be *independent* if the occurrence of one event does not affect the probability of the occurrence of the other event. In other words, *A* and *B* are independent events if

$$\text{either} \quad P(A \mid B) = P(A) \quad \text{or} \quad P(B \mid A) = P(B)$$

A and *B* are dependent events if

$$P(A \mid B) \neq P(A) \text{ or } P(B \mid A) \neq P(B).$$

4. Give one example for the mutually exclusive event, independent event, and dependent event, respectively.

- Mutually exclusive event-house ownership

- Independent event- profession
- Dependent event- is experience on income

Complementary Events :-

The complement of event *A*, denoted by *A* and read as “*A* bar” or “*A* complement,” is the event that includes all the outcomes for an experiment that are not in *A.*

5. Give one example for the complementary event.
    - Marital status is a complimentary event, which has two outcomes.

Intersection of Events :-

Let *A* and *B* be two events defined in a sample space. The intersection of *A* and *B* represents the collection of all outcomes that are common to both *A* and *B* and is denoted by A and B or A ∩ B or AB.

6. Give one example for the intersection of event.
    - State and city

Union of Events :-

Let *A* and *B* be two events defined in a sample space. The **union of events** *A* and *B* is the collection of all outcomes that belong either to *A* or to *B* or to both *A* and *B* and is denoted by (*A* or *B*) or *A* ∪ *B.*

7. Give one example for the union of events.
    - There is no union of events in the dataset.

Week 8: Chapter 7: Sampling distributions

Sampling distributions: Standard deviation of a Sample mean

	Income Total	Ages	Number of Years Employed	Flag Mobile	Flag work Phone	Flag Phone	Flag Email	Begin Months	Target
count	537667	537667	537667	537667	537667	537667	537667	537667	537667
Mean	197117.13	41.13	7.57	1.0	0.28	0.30	0.1	19.31	0.00
Std	104138.96	9.36	6.56	0.0	0.45	0.46	0.3	14.04	0.06
min	27000.00	20.52	0.05	1.0	0.00	0.00	0.0	0.00	0.00
25%	135000.00	33.53	2.88	1.0	0.00	0.00	0.0	8.00	0.00
50%	180000.00	40.51	5.88	1.0	0.00	0.00	0.0	17.00	0.00
75%	229500.00	48.20	10.03	1.0	1.00	1.00	0.0	29.00	0.00
max	1575000.00	67.43	43.05	1.00	1.00	1.00	1.00	60.00	1.00

Page 24, please create an example similar to Example 7-2 using your data and solve it.

The mean of Total Income per year is $\bar{x}$ =$197117.13

standard deviation is σ $104138.96

The count of total people N 537667

Let $\bar{x}$ be the mean of Total Income per year a random sample of certain people selected from the data. Find the mean and standard deviation of x for a sample size of people:

a) n= 20000

When $\frac{n}{N} \leq 0.05 = 0.0371$, the standard deviation of the sampling distribution of $\bar{x}$ is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. = 104138.96/141.42= 738.30

b) n= 10000

When $\frac{n}{N} \leq 0.05 = 0.0186$, the standard deviation of the sampling distribution of $\bar{x}$ is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. = 104138.96/100= 1041.38

c) 500000

When $\frac{n}{N} > 0.05 = 0.92$, $\sigma_x = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$.= 147.27*0.264681 = 38.98

Sampling distributions: Lead time distribution

Page 34, please create an example similar to Example 7-5 using your data and solve it

Assume that the number of working years or years employed of all people are approximately normally distributed with a mean of 7.57 years and a standard deviation of 6.56 years. Find the probability that the mean , *x*, of a random sample of 500 people of the data between 3 and 6 years.

- μ_x = 7.57 years, as n/N< 0.05
- The standard deviation of 5000 sample is:
  $\sigma_x = \sigma / \sqrt{n}$ = 7.57/ √500 = 7.57/7.07 =1.070

P (3< $\bar{x}$ < 6)

Z= $(\bar{x} - \mu)/ \sigma_x$ = (3 – 7.57)/1.070 = -4.27

Z= $(\bar{x} - \mu)/ \sigma_x$ = (6 – 7.57)/1.070= -1.467

Steps:

For $\bar{x}$=3, Z score = $\frac{x - \mu}{\sigma}$

= $\frac{3 - 7.57}{1.07}$

= -4.27103

P-value from Z-Table:

P(x<3) = 0.00001

P(x>3) = 1 - P(x<3) = 0.99999

P(3<x<7.57) = 0.5 - P(x<3) = 0.49999

For $\bar{x}$=6, Z score = $\frac{x - \mu}{\sigma}$

= $\frac{6 - 7.57}{1.07}$

= -1.46729

P-value from Z-Table:

P(x<6) = 0.071149

P(x>6) = 1 - P(x<6) = 0.92885

P(6<x<7.57) = 0.5 - P(x<6) = 0.42885

Looking at the probability chart, the result is:

P(3< $\bar{x}$ <6)=P(-4.27<z<-1.467)

= 0.49999-0.42885

=0.07114

Sampling distributions: Central limit theorem for sample proportion:

Page 47, please create an example similar to Example 7-9 using your data and solve.

According to the data set the number of people that have flag work phone is 151415 and percent of people who have flag work phone is:

(Number of people that have flag work phone/ Total number of people) *100:
= (151415 ÷ 537667)*100 = 28.16%

Assume that this result is true for the current population of data set. Let $\hat{p}$ be the proportion of people that have flag work phone in a random sample of 2000. Find the mean and standard deviation of $\hat{p}$ and describe the shape of its sampling distribution.

Let $\hat{p}$ be the proportion of people that have flag work phone: 0.281

$q = 1 - p = 1 – 0.281 = 0.719$

The mean of the sampling distribution of $\hat{p}$ is

$\mu_p = p = 0.281$

The standard deviation of $\hat{p}$ is:

$\sigma_p = \sqrt{(pq/n)}$
= √(0.281*0.719/2000)

= √(0.221709/2000)

= √0.00011085

= 0.105

As $np = 2000(0.281) = 562$

As $nq = 2000(0.719) = 1438$

Both are greater than 5, the central limit theorem is applied, so the sampling distribution of $\hat{p}$ is approximately normal with mean 0.072 and standard deviation of 0.105

Sampling distributions: Application of the sampling distribution of $\hat{p}$:

Page 52, please create an example similar to Example 7-10 using your data and solve it.

According to the data set the number of people that have flag phone is 160705 and percent of people who have flag phone is:

(Number of people that have flag work phone/ Total number of people) *100:
= (160705 ÷ 537667)*100 = 29.88%

Suppose that this result is true for the dataset. Let $\hat{p}$ be the proportion in a random sample of 2000 people that have flag work phone. Find the probability that 30% to 32% of adults in this sample will hold this opinion.

= 2000, $p = 0.2988$, and $q = 1 - p = 1 - 0.298 = 0.702$, where p is the proportion people that have flag work phone

The mean of the sample proportion $\hat{p}$ is $\mu_p = p = 0.298$

The standard deviation of $\hat{p}$ is $\sigma_p = \sqrt{(pq/n)} = \sqrt{(0.298*0.702/2000)}$
$= \sqrt{0.209196/2000} = 0.01022$

As $np = 2000(0.298) = 596$

As $nq = 2000(0.702) = 1404$

As both are greater than 5, the central limit theorem could be applied to infers that the data is distributed approximately normal.

$= p\,(0.3 < \hat{p} < 0.32)$

$z = (\bar{p} - p)/\sigma_p = (0.3-0.298)/0.01022$

- Let $\bar{p} = x$, for $x = 0.3$, Z score $= \frac{x - \mu}{\sigma}$

$= \frac{0.3 - 0.298}{0.01022}$

$=0.19569$

- P-value from Z-Table:
  P(x<0.3) = 0.57758
  P(x>0.3) = 1 - P(x<0.3) = 0.42242
  P(0.298<x<0.3) = P(x<0.3) - 0.5 = 0.077575

$z = (\bar{p} - p)/\sigma_p$ = (0.32-0.298)/0.01022

- Let $\bar{p}$= x, for x = 0.32, Z score = $\frac{x - \mu}{\sigma}$

$= \frac{0.32 - 0.298}{0.01022}$

=2.15264

- P-value from Z-Table:
  P(x<0.32) = 0.98433
  P(x>0.32) = 1 - P(x<0.32) = 0.015673
  P(0.298<x<0.32) = P(x<0.32) - 0.5 = 0.48433

Finding in the z table, the probability is 0.0006
P (z<0.32)- P(z>0.3) = 0.48433-0.077575= 0.406755 =0.41

Week 9 Chapter 8: Estimations of the mean and proportion:

Page 17, please create an example similar to Example 8-1 using your data and solve *it. Please refer to page 22 and interpret your confidence interval.*

Table 8.1 ***z*** **Values for Commonly Used Confidence Levels**

Confidence Level	Areas to Look for in Table IV	*z* Value
90%	.0500 and .9500	1.64 or 1.65
95%	.0250 and .9750	1.96
96%	.0200 and .9800	2.05
97%	.0150 and .9850	2.17
98%	.0100 and .9900	2.33
99%	.0050 and .9950	2.57 or 2.58

In the data set took a random sample of 32 people of their number of years employed. We found out mean of their number of years employed for this sample is 8.589 years and standard deviation is 6.56.

Construct a 90% confidence interval for the mean lead time.

- Here, n> 30, hence we can use the normal distribution.

- From the given information, n = 32, $\bar{x}$ = 8.589, and σ = 6.56

➢ $\sigma_x = \sigma/\sqrt{n}$
➢ $= 6.56/\sqrt{32} = 1.16$

We know that z = -1.65 and z = 1.65

The 90% confidence interval for μ is $\bar{x} \pm z\sigma_x = 8.589 \pm 1.65*1.16$

➢ $= 8.589 \pm 1.194$
➢ = 7.429 to 10.503

Therefore, 90% confidence interval of number of years employed is 7.429 to 10.503 years for 32 people sample data.

(a) The point estimate of the proportion of number of years employed is equal to 8.589; that is. Point estimate of p= $\hat{p}$ = 0.80

(b) The confidence level is 90% or 0.9.

From the given information, n = 32, $\bar{x}$ = 8.589, and σ = 6.56

➢ $\sigma_x = \sigma/\sqrt{n}$
➢ $= 6.56/\sqrt{32} = 1.16$

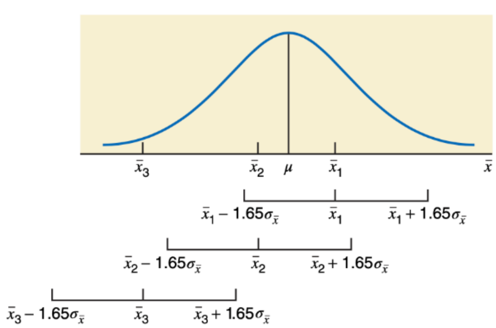
We know that z = -1.65 and z = 1.65

The 90% confidence interval for μ is $\bar{x} \pm z\sigma_x = 8.589 \pm 1.65*1.16$

➢ $= 8.589 \pm 1.194$
➢ = 7.429 to 10.503

Therefore, 90% confidence interval of number of years employed is 7.429 to 10.503 years for 32 people sample data.

Figure 8.4 Confidence intervals.



[1]Note that there is no apparent reason for choosing .0495 and .9505 and not choosing .0505 and .9495 in Table IV. If we choose .0505 and .9495, the z values will be −1.64 and 1.64. An alternative is to use the average of 1.64 and 1.65, which is 1.645, which we will not do in this text.

♦ If we take all possible samples 32 of number of years employed and construct a 90% confidence interval for μ around each sample mean, we can expect that 90% of these intervals will include μ and 10% will not.

Conclusion

Limitations:

- The dataset only includes 537667 records of Indian consumers, there might be differences between recorded consumers and non- recorded consumers.
- Model can be improved with more data and computational resources.

Key findings:

- Learn about the Costumers behaviors and characteristics for Credit Card
- Costumers with property can also have default risk.
- Model can be served as an aid to human decision.
-

Recommendations:

- The banks or financial institutions may use the scorecard to develop a detailed ranking system for filtering of credit card applications.
- Some additional parameters such as tightening of FICO score, restricting the line of credit for flagged off customers etc. can be implemented.

References:

- **The data is taken from Kaggle datasets which is a web service platform for data. https://www.kaggle.com/laotse/credit-card-approval**
- **https://core.ac.uk/download/pdf/77009321.pdf**
- **https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1024.2660&rep=rep1&type=pdf**
- Pak Shun HON and Tony BELLOTTI. (n.d.). *Models and forecasts of credit card balance*. Retrieved December 14, 2021, from https://core.ac.uk/download/pdf/77009321.pdf.

Results , Key figures and Tables

Exploratory Data Analysis

- The dataset comprises of 537667 observations and 19 characteristics. Amongst all the variables, Target is the dependent variable and the rest 18 are independent variables.
- Data types of each variable in the dataset

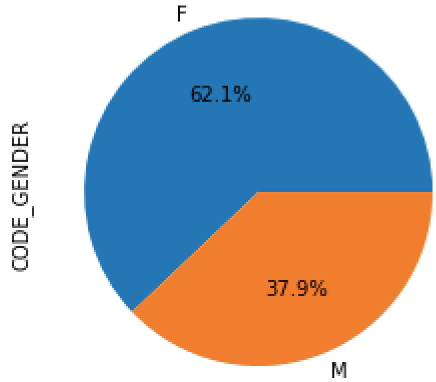
Variables	**Types**
ID	Int64
CODE_GENDER	Object
FLAG_OWN_CAR	Object
FLAG_OWN_REALTY	Object
CNT_CHILDREN	Object
AMT_INCOME_TOTAL	Float64
NAME_EDUCATION_TYPE	Object
NAME_FAMILY_STATUS	Object
NAME_HOUSING_TYPE	Object

AGES	Float64
NUMBER_OF_YEARS_EMPLOYED	Float64
FLAG_MOBIL	Int64
FALG_WORK_PHONE	Int64
FLAG_PHONE	Int64
FLAG_EMAIL	Int64
JOB	Object
BEGIN_MONTHS	Int64
STATUS	Object
TARGET	Int64

- Number of missing values

Variables	**Number of Missing Values**
ID	0
CODE_GENDER	0
FLAG_OWN_CAR	0
FLAG_OWN_REALTY	0
CNT_CHILDREN	0
AMT_INCOME_TOTAL	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
NAME_HOUSING_TYPE	0
AGES	0
NUMBER_OF_YEARS_EMPLOYED	0
FLAG_MOBIL	0
FALG_WORK_PHONE	0
FLAG_PHONE	0
FLAG_EMAIL	0
JOB	0
BEGIN_MONTHS	0
STATUS	0
TARGET	0

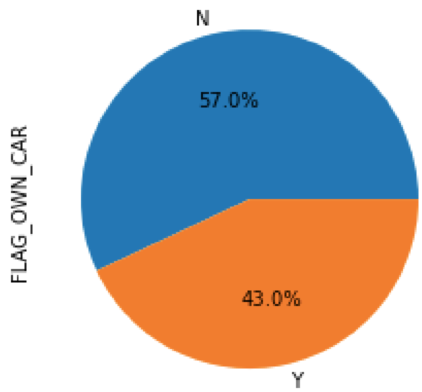
- Number of Males and Females:



	F	M
CODE_GENDER	333832	203835

See Appendix 14

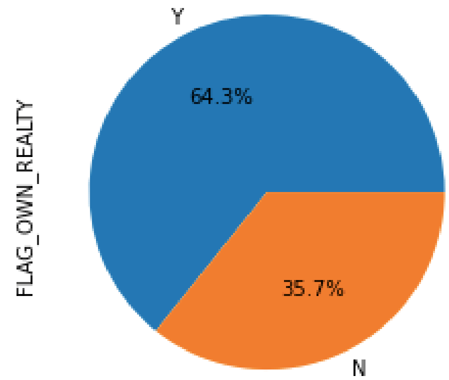
- Number of Car Ownership:



	N	Y
FLAG_OWN_CAR	306207	231460

See Appendix 15

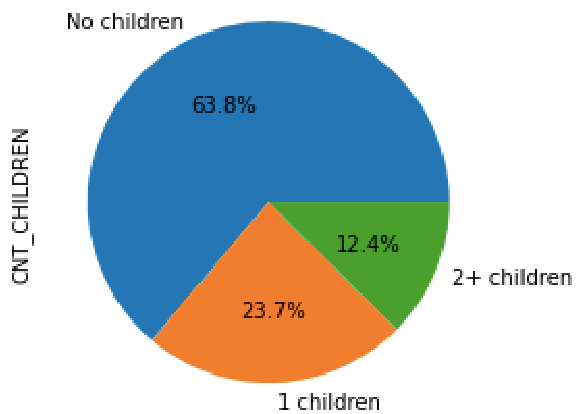
- Number of Properties Ownership:



	Y	N
FLAG_OWN_REALTY	345471	192196

See Appendix 16

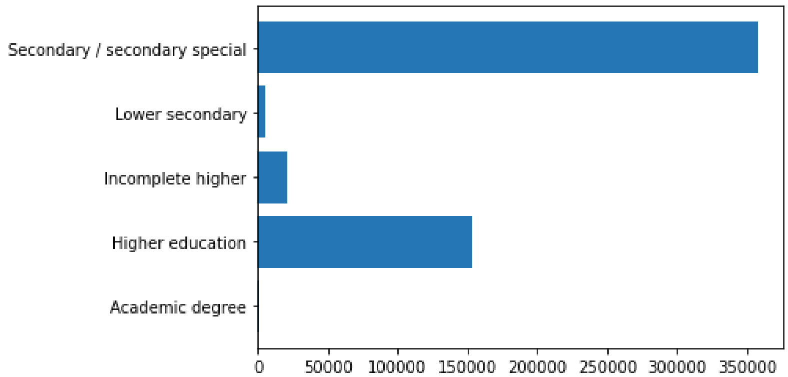
- Number of Children Count:



	No children	1 children	2+ children
CNT_CHILDREN	343151	127695	66821

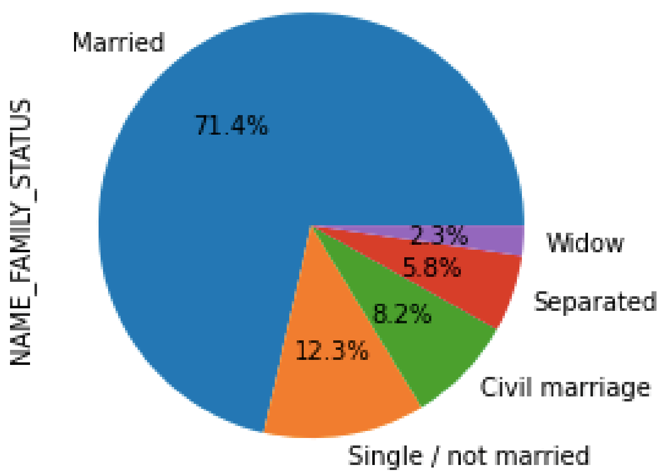
See Appendix 17

- Types of Education:



See Appendix 18

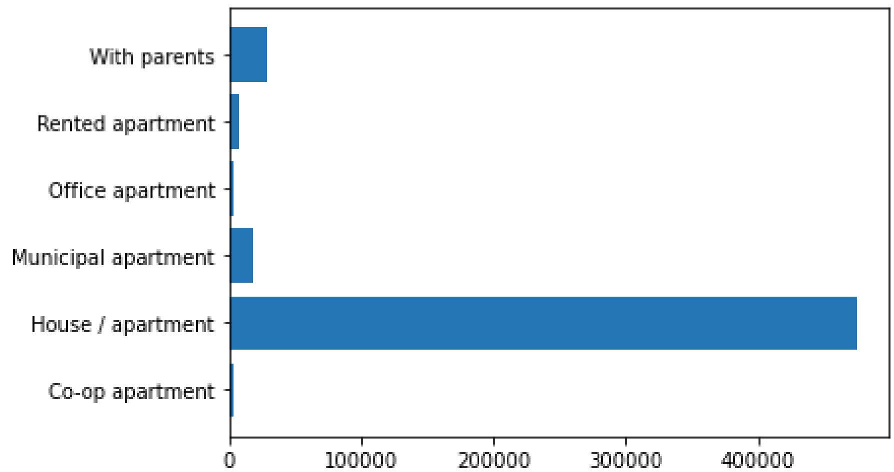
- Type of family status:



	Married	Single / not married	Civil marriage	Separated	Widow
NAME_FAMILY_STATUS	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

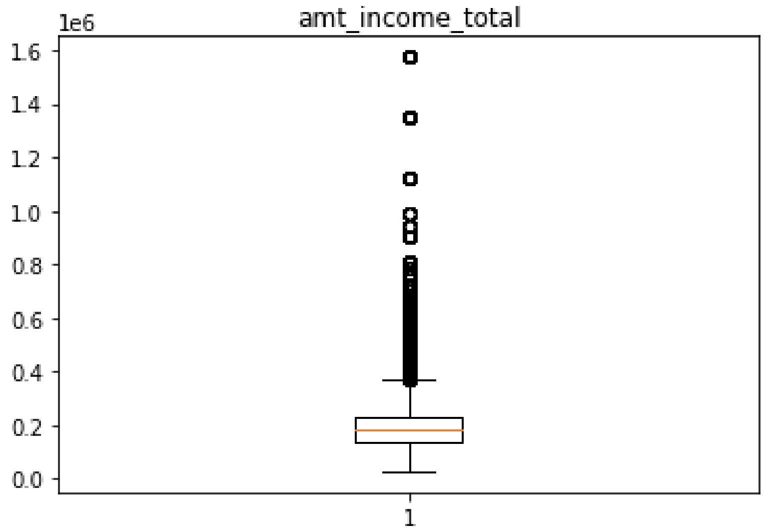
See Appendix 19

- Type of Housing Type:



See Appendix 20

- After doing the exploratory, we figured out that our dataset does not have any missing value. Also, we found some key findings:
    - The number of females is larger than the number of males
    - Most observations in our dataset are living in the house/apartment
    - The number of people who own cars is higher than the number of people who do not own car
    - The number of people who own property is higher than the number of people who do not own property
    - The percentage of people who do not have any children is the highest with 63.8%
    - The level of education falls mostly in Secondary/Secondary special which means that these people are mostly low educated
    - Most of them are married
- Most of the mean of the variables in the dataset is larger than the median, so their histogram is right-skewed distribution
- There is notably a large different between the 75^{th} %tile and max values of the variable Annual Income and number of years working, which means that in this variable, we have a lot of outliers.



See Appendix 21

See Appendix 22

Regression

Appendix

Appendix 1

```
import random
df.sample(n=5)
```

Appendix 2

```
systematic_sample_df = df.iloc[::5]
systematic_sample_df.head()
```

Appendix 3

```
customers = {'ID':['5036943','5023870','5021648','5112626','5089398'],'CODE_GENDER':['F','F','F','F','F'],'CNT_CHILD
customers
```

```
import pandas as pd
pd.DataFrame.from_dict(customers)
```

Appendix 4

```
import pandas as pd
import numpy as ny

customers1 = {'ID':np.arange(5000000,5999999)}
customers1
pd.DataFrame.from_dict(customers1)
```

Appendix 5

```
pd.crosstab(index=df['JOB'],columns='count')
```

Appendix 6

```
relative_frequencies = df['JOB'].value_counts(normalize=True)
relative_frequencies
```

Appendix 7

```python
round(df1.describe(), 2)
```

	AMT_INCOME_TOTAL	AGES	NUMBER_OF_DAYS_EMPLOYED	FLAG_MOBIL	FLAG_WORK_PHONE	FLAG_PHONE	FLAG_EMAIL	BEGIN_MONTHS	
count	537667.00	537667.00	537667.00	537667.0	537667.00	537667.00	537667.0	537667.00	5
mean	197117.13	41.13	7.57	1.0	0.28	0.30	0.1	19.31	
std	104138.96	9.36	6.56	0.0	0.45	0.46	0.3	14.04	
min	27000.00	20.52	0.05	1.0	0.00	0.00	0.0	0.00	
25%	135000.00	33.53	2.88	1.0	0.00	0.00	0.0	8.00	
50%	180000.00	40.51	5.88	1.0	0.00	0.00	0.0	17.00	
75%	229500.00	48.20	10.03	1.0	1.00	1.00	0.0	29.00	
max	1575000.00	67.43	43.05	1.0	1.00	1.00	1.0	60.00	

Appendix 8

```python
#Trimmed mean
```

```python
stats.trim_mean(df1['AMT_INCOME_TOTAL'],0.1)
```

```
183963.7261383054
```

```python
stats.trim_mean(df1['NUMBER_OF_DAYS_EMPLOYED'],0.1)
```

```
6.521652109221524
```

```python
stats.trim_mean(df1['BEGIN_MONTHS'],0.1)
```

```
18.150543434038152
```

Appendix 9

```
#Mode
```

```
df1['AMT_INCOME_TOTAL'].mode()
```

```
0    135000.0
dtype: float64
```

```
df1['CODE_GENDER'].mode()
```

```
0    F
dtype: object
```

```
df1['FLAG_OWN_CAR'].mode()
```

```
0    N
dtype: object
```

```
df1['FLAG_OWN_REALTY'].mode()
```

```
0    Y
dtype: object
```

```
df1['CNT_CHILDREN'].mode()
```

```
0    No children
dtype: object
```

```
df1['NAME_EDUCATION_TYPE'].mode()
```

```
0    Secondary / secondary special
dtype: object
```

```
df1['NAME_FAMILY_STATUS'].mode()
```

```
0    Married
dtype: object
```

```
df1['JOB'].mode()
```

```
0    Laborers
dtype: object
```

```
df1['BEGIN_MONTHS'].mode()
```

```
0    1
dtype: int64
```

```
df1['STATUS'].mode()
```

```
0    C
dtype: object
```

```
df1['TARGET'].mode()
```

```
0    0
dtype: int64
```

```
df1['NAME_HOUSING_TYPE'].mode()
```

```
0    House / apartment
dtype: object
```

```
df1['AGES'].mode()
```

```
0    34.73
dtype: float64
```

```
df1['NUMBER_OF_DAYS_EMPLOYED'].mode()
```

```
0    1.18
dtype: float64
```

```
df1['FLAG_MOBIL'].mode()
```

```
0    1
dtype: int64
```

```
df1['FLAG_WORK_PHONE'].mode()
```

```
0    0
dtype: int64
```

```
df1['FLAG_PHONE'].mode()
```

```
0    0
dtype: int64
```

```
df1['FLAG_EMAIL'].mode()
```

```
0    0
dtype: int64
```

Appendix 10

```
#Calculate weighted mean of Annual income and Ages
```

```
a= df1['AMT_INCOME_TOTAL'].iloc[:10]
b = df1['AGES'].iloc[:10]
print(list(a))
print(list(b))
```

```
[270000.0, 81000.0, 270000.0, 112500.0, 139500.0, 144000.0, 180000.0, 405000.0, 135000.0, 270000.0]
[36.32, 48.98, 53.64, 41.39, 47.35, 42.18, 30.62, 51.11, 46.76, 45.52]
```

```
def weighted_mean(a,b):
    return round(sum([a[i]*b[i] for i in range(len(a))])/sum(b),2)
weighted_mean(a,b)
```

```
203677.54
```

Appendix 11

```
#Calculate the percentile of 20 observations using one variable
```

```
c = df1['AMT_INCOME_TOTAL'].iloc[:20]
print(c)
```

```
0     270000.0
1      81000.0
2     270000.0
3     112500.0
4     139500.0
5     144000.0
6     180000.0
7     405000.0
8     135000.0
9     270000.0
10     99000.0
11    103500.0
12    225000.0
13    171000.0
14    135000.0
15    270000.0
16    225000.0
17    202500.0
18    135000.0
19     67500.0
Name: AMT_INCOME_TOTAL, dtype: float64
```

```
print(np.percentile(c,25))
```

```
129375.0
```

```
print(np.percentile(c,50))
```

```
157500.0
```

```
print(np.percentile(c,75))
```

```
236250.0
```

Appendix 12

```
#Calculate the percentile rank of 20 observations using one variable
```

```
df1["Percentile Rank"] = df1.AMT_INCOME_TOTAL.rank(pct=True)
df1.head(20)
```

Appendix 13

```
#Create the box and whisker plot of 30 observations using one variable
```

```
d = df1['AMT_INCOME_TOTAL'].iloc[:30]
plt.boxplot(d)
plt.show()
```

Appendix 14

```
df['CODE_GENDER'].value_counts().plot(kind='pie',table=True, autopct='%1.1f%%')
```

Appendix 15

```
df['FLAG_OWN_CAR'].value_counts().plot(kind='pie',table=True, autopct='%1.1f%%')
```

Appendix 16

```
df['FLAG_OWN_REALTY'].value_counts().plot(kind='pie',table=True, autopct='%1.1f%%')
```

Appendix 17

```
df['CNT_CHILDREN'].value_counts().plot(kind='pie',table=True, autopct='%1.1f%%')
```

Appendix 18

```
default_housing_type = df[['NAME_EDUCATION_TYPE','TARGET']]
default_housing_type.head()
```

```
df_tmp = default_housing_type.groupby(['NAME_EDUCATION_TYPE']).count()

df_tmp = df_tmp.reset_index()
df_tmp
```

```
plt.barh(df_tmp['NAME_EDUCATION_TYPE'], df_tmp['TARGET'])
plt.show()
```

Appendix 19

```
df['NAME_FAMILY_STATUS'].value_counts().plot(kind='pie',table=True, autopct='%1.1f%%')
```

Appendix 20

```
default_housing_type = df[['NAME_HOUSING_TYPE','TARGET']]
default_housing_type.head()
```

```
df_tmp = default_housing_type.groupby(['NAME_HOUSING_TYPE']).count()

df_tmp = df_tmp.reset_index()
df_tmp
```

```
plt.barh(df_tmp['NAME_HOUSING_TYPE'], df_tmp['TARGET'])
plt.show()
```

Appendix 21

```
plt.boxplot(df['AMT_INCOME_TOTAL'])
plt.title('amt_income_total')
plt.show()
```

Appendix 22

```
plt.boxplot(df['NUMBER_OF_DAYS_EMPLOYED'])
plt.title('number of years employed')
plt.show()
```

```
from sklearn import metrics
meanAbErr = metrics.mean_absolute_error(y_test, y_pred_mlr)
meanSqErr = metrics.mean_squared_error(y_test, y_pred_mlr)
rootMeanSqErr = np.sqrt(metrics.mean_squared_error(y_test, y_pred_mlr))
print('R squared: {:.2f}'.format(mlr.score(x,y)*100))
print('Mean Absolute Error:', meanAbErr)
print('Mean Square Error:', meanSqErr)
print('Root Mean Square Error:', rootMeanSqErr)
```

```
mlr_diff = pd.DataFrame({'Actual value': y_test, 'Predicted value': y_pred_mlr})
mlr_diff.head()
```

```
y_pred_mlr= mlr.predict(x_test)
#Predicted values
print("Prediction for test set: {}".format(y_pred_mlr))
```

```
print("Intercept: ", mlr.intercept_)
print("Coefficients:")
list(zip(x, mlr.coef_))
```

```
In [21]: dataset = df
         dataset = dataset.iloc[:750,:]
         dataset = dataset.dropna()
```

```
In [22]: import seaborn as seabornInstance
         from sklearn.model_selection import train_test_split
         from sklearn.linear_model import LinearRegression
         from sklearn import metrics
```

```
In [24]: x = dataset [['ID','AMT_INCOME_TOTAL','AGES','NUMBER_OF_DAYS_EMPLOYED','FLAG_MOBIL','FLAG_WORK_PHONE','FLAG_PHONE','FLAG_EMAIL','BEGIN_MONTHS']]
         y = dataset['TARGET']
```

```
In [25]: x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 10 )
```

```
In [26]: mlr = LinearRegression()
         mlr.fit(x_train, y_train)
```

```
In [20]: import pandas as pd
         import numpy as np
         import seaborn as sns
         import matplotlib.pyplot as plt
         from scipy import stats
         raw = r'/Users/charesh/Downloads/credit_card_approval_Process.csv'
         df=pd.read_csv(raw)
         df.head()
```