Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:05

Topics:

0. ANOVA

1. Kruskal-Wallis [H] Test → when ANOVA breaks down — Non-param

2. Plot : don't just see numbers

3. Biased estimate → why we use (n-1) for std-dev

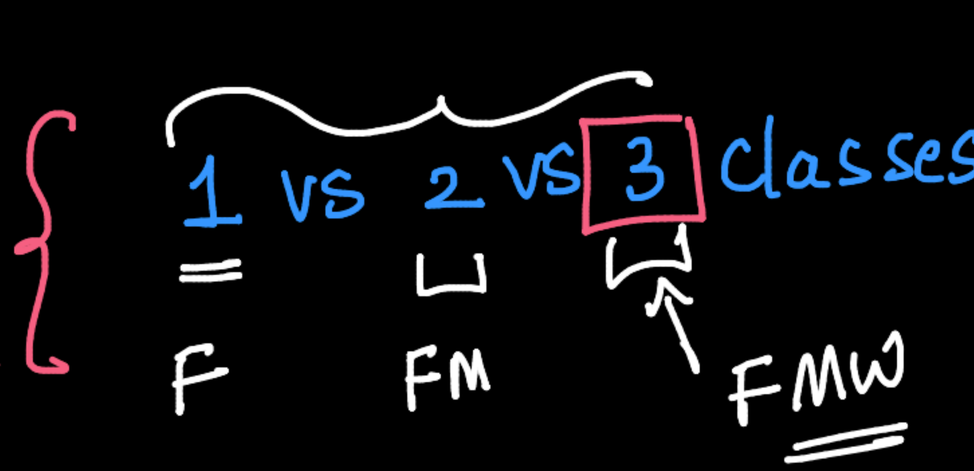
4. Case-study: Loan granting → Feature Engg.. using everything we know

5. Handle missing values

Post-read ↓

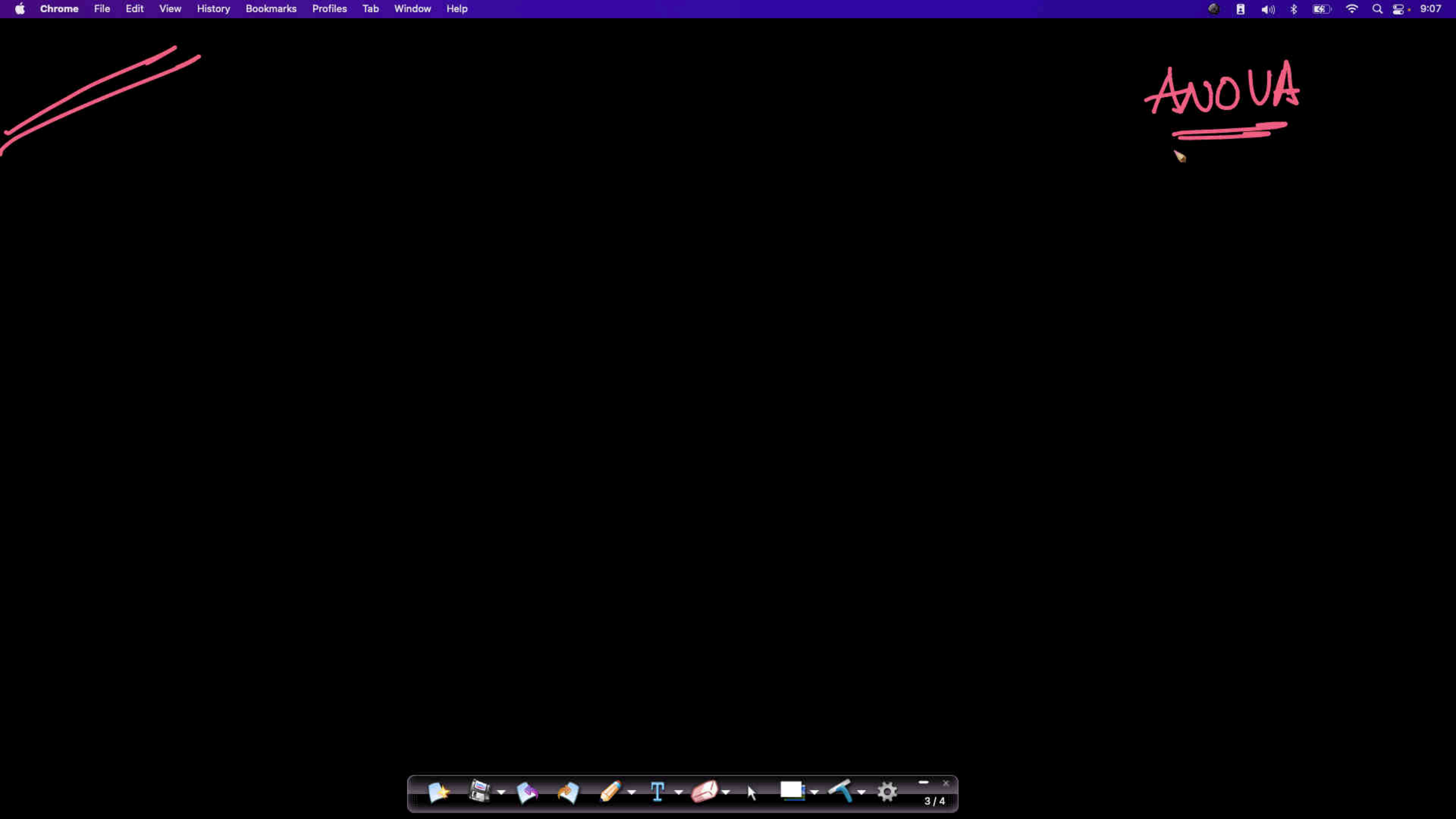
Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:06

Ops:



↳ optional problem solving → pick least freq-answer

↳ whatsapp → Tag Ajay + Assignments team


Chrome File Edit View History Bookmarks Profiles Tab Window Help
9:07
ANOVA
3 / 4

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:07
10.2 - Hypothesis Testing | Kruskal–Wallis one-way a | Anscombe's quartet - Wik | Bias of an estimator - Wik | EDA_FE.ipynb - Colaborat | Markov's inequality - Wiki | Chebyshev's inequality - | New Tab
online.stat.psu.edu/stat200/lesson/10/10.2 Update

A one-way between groups ANOVA is used to compare the means of more than two independent groups. A one-way between groups ANOVA comparing just two groups will give you the same results at the independent t test that you conducted in Lesson 8. We will use the five step hypothesis testing procedure again in this lesson.

1. Check assumptions and write hypotheses

The assumptions for a one-way between groups ANOVA are:

1. Samples are independent
2. The response variable is approximately normally distributed for each group or all group sample sizes are at least 30
3. The population variances are equal across responses for the group levels (if the largest sample standard deviation divided by the smallest sample standard deviation is **not** greater than **two**, then assume that the population variances are equal)

Given that you are comparing k independent groups, the null and alternative hypotheses are:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
H_a : Not all $\mu_.$ are equal

In other words, the null hypothesis is that at all of the groups' population means are equal. The alternative is that they are not all equal; there are at least two population means that are not equal to one another.

2. Calculate the test statistic

ANOVA uses an F test statistic. Hand calculations for ANOVAs require many steps. In this class, you will be working primarily with Minitab outputs.

Conceptually, the F statistic is a ratio: $F = \frac{Between\ groups\ variability}{Within\ groups\ variability}$. Numerically this translates to $F = \frac{MS_{Between}}{MS_{Within}}$. In other words how much do individuals in different groups vary from one another over how much to individuals within groups vary from one another.

Statistical software will compute the F ratio fo VA

Resources

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:08
10.2 - Hypothesis Testing | Kruskal–Wallis one-way a | Anscombe's quartet - Wik | Bias of an estimator - Wik | EDA_FE.ipynb - Colaborat | Markov's inequality - Wiki | Chebyshev's inequality - | New Tab
online.stat.psu.edu/stat200/lesson/10/10.2 Update

A one-way between groups ANOVA is used to compare the means of more than two independent groups. A one-way between groups ANOVA comparing just two groups will give you the same results at the independent t test that you conducted in Lesson 8. We will use the five step hypothesis testing procedure again in this lesson.

1. Check assumptions and write hypotheses

box-cox

QQ / KS / AD

rec-times

The assumptions for a one-way between groups ANOVA are:

1. Samples are independent
2. The response variable is approximately normally distributed for each group or all group sample sizes are at least 30
3. The population variances are equal across responses for the group levels (if the largest sample standard deviation divided by the smallest sample standard deviation is **not** greater than **two**, then assume that the population variances are equal)

Given that you are comparing k independent groups, the null and alternative hypotheses are:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
H_a : Not all $\mu_.$ are equal

In other words, the null hypothesis is that at all of the groups' population means are equal. The alternative is that they are not all equal; there are at least two population means that are not equal to one another.

2. Calculate the test statistic

ANOVA uses an F test statistic. Hand calculations for ANOVAs require many steps. In this class, you will be working primarily with Minitab outputs.

Conceptually, the F statistic is a ratio: $F = \frac{Between\ groups\ variability}{Within\ groups\ variability}$. Numerically this translates to $F = \frac{MS_{Between}}{MS_{Within}}$. In other words how much do individuals in different groups vary from one another over how much to individuals within groups vary from one another.

Statistical software will compute the F ratio fo VA

Resources

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:09
10.2 - Hypothesis Testing | Kruskal–Wallis one-way a | Anscombe's quartet - Wik | Bias of an estimator - Wik | EDA_FE.ipynb - Colaborat | Markov's inequality - Wiki | Chebyshev's inequality - | New Tab
online.stat.psu.edu/stat200/lesson/10/10.2 Update

A one-way between groups ANOVA is used to compare the means of more than two independent groups. A one-way between groups ANOVA comparing just two groups will give you the same results at the independent t test that you conducted in Lesson 8. We will use the five step hypothesis testing procedure again in this lesson.

1. Check assumptions and write hypotheses

The assumptions for a one-way between groups ANOVA are:

1. Samples are independent
2. The response variable is approximately normally distributed for each group or all group sample sizes are at least 30
3. The population variances are equal across responses for the group levels (if the largest sample standard deviation divided by the smallest sample standard deviation is **not** greater than **two**, then assume that the population variances are equal)

$s_1 = 10$ $s_2 = 20$ $s_3 = 18$

Given that you are comparing k independent groups, the null and alternative hypotheses are:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
H_a : Not all $\mu_.$ are equal

$s_1 = 10$ $s_2 = 100$

In other words, the null hypothesis is that at all of the groups' population means are equal. The alternative is that they are not all equal; there are at least two population means that are not equal to one another.

2. Calculate the test statistic

ANOVA uses an F test statistic. Hand calculations for ANOVAs require many steps. In this class, you will be working primarily with Minitab outputs.

Conceptually, the F statistic is a ratio: $F = \frac{Between\ groups\ variability}{Within\ groups\ variability}$. Numerically this translates to $F = \frac{MS_{Between}}{MS_{Within}}$. In other words how much do individuals in different groups vary from one another over how much to individuals within groups vary from one another.

Statistical software will compute the F ratio fo VA

Resources

Datasets Glossary Formulas
Contact Help & Support
Minitab Quick Guide

6 / 7

$$\begin{cases} n = m \text{ per gp} \\ \times \\ k \text{ gps} \end{cases}$$

$$T_{\text{Anova}} = f = \frac{\frac{1}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 \times m}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2}$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help
10.2 - Hypothesis Testing | Kruskal–Wallis one-way a | Anscombe's quartet - Wi | Bias of an estimator - Wik | EDA_FE.ipynb - Colabora | Markov's inequality - Wiki | Chebyshev's inequality - | New Tab
online.stat.psu.edu/stat200/lesson/10/10.2
Update

table. The ANOVA source table will give you information about the variability between groups and within groups. **The table below gives you all of the formulas, but you will not be responsible for performing these calculations by hand. Minitab will do all of these calculations for you and provide you with the full ANOVA source table.**

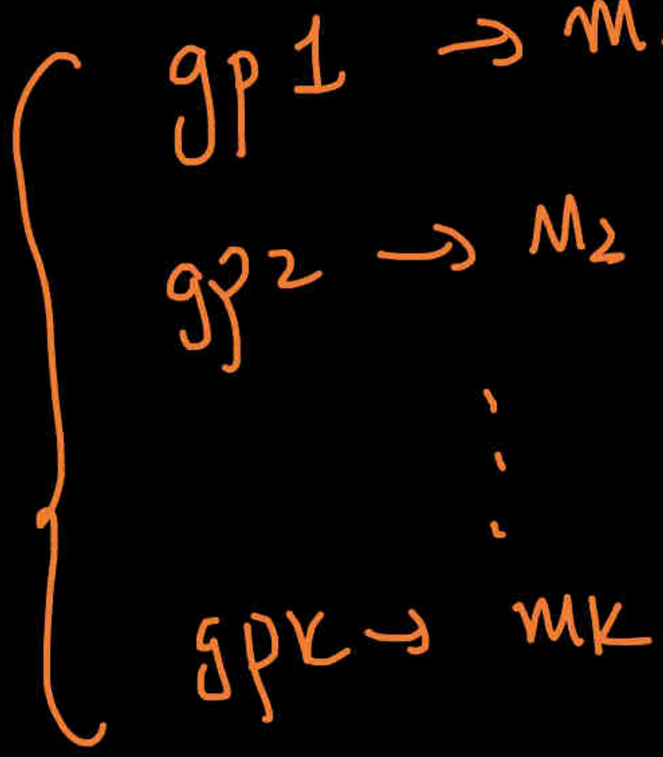
m = # obs per group

Source	df	SS	MS	F	p
Between Groups (Factor)	$k-1$	$\sum_k n_k(\overline{x}_k - \overline{x}_.)^2$	$\frac{SS_{Between}}{df_{Between}}$	$\frac{MS_{Between}}{MS_{Within}}$	Area to the right of $F_{k-1,\, n-k}$
Within Groups (Error)	$n-k$	$\sum_k \sum_i (x_{ik} - \overline{x}_k)^2$	$\frac{SS_{Within}}{df_{Within}}$		
Total	$n-1$	$\sum_k \sum_i (x_{ik} - \overline{x}_.)^2$			

Legend

k	Number of groups
n	Total sample size (all groups combined)
n_k	Sample size of group k
$\overline{x}_k$	Sample mean of group k
$\overline{x}_.$	Grand mean (i.e., mean for all groups combined)
SS	Sum of squares
MS	Mean square

8 / 8



$n = m_1 + m_2 + \cdots + m_k$

→ scipy.stats

→ statsmodels →

Chrome File Edit View History Bookmarks Profiles Tab Window Help
10.2 - Hypothesis Testing | Kruskal–Wallis one-way a | Anscombe's quartet - Wi | Bias of an estimator - Wik | EDA_FE.ipynb - Colaborat | Markov's inequality - Wiki | Chebyshev's inequality - | New Tab
online.stat.psu.edu/stat200/lesson/10/10.2
Update

groups. A one-way between groups ANOVA comparing just two groups will give you the same results at the independent t test that you conducted in Lesson 8. We will use the five step hypothesis testing procedure again in this lesson.

1. Check assumptions and write hypotheses

The assumptions for a one-way between groups ANOVA are:

1. Samples are independent
2. The response variable is approximately normally distributed for each group or all group sample sizes are at least 30
3. The population variances are equal across responses for the group levels (if the largest sample standard deviation divided by the smallest sample standard deviation is **not** greater than **two**, then assume that the population variances are equal)

Given that you are comparing k independent groups, the null and alternative hypotheses are:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
H_a : Not all $\mu_.$ are equal

In other words, the null hypothesis is that at all of the groups' population means are equal. The alternative is that they are not all equal; there are at least two population means that are not equal to one another.

2. Calculate the test statistic

ANOVA uses an F test statistic. Hand ... you will

Resources

Datasets Glossary

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:17
10.2 - Hypothesis Testing | ST… Kruskal–Wallis one-way analy… Anscombe's quartet - Wikiped… Bias of an estimator - Wikiped… EDA_FE.ipynb - Colaboratory Markov's inequality - Wikiped… Chebyshev's inequality - Wiki…
en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance
Update

Kruskal–Wallis one-way analysis of variance

From Wikipedia, the free encyclopedia

The **Kruskal–Wallis test** by ranks, **Kruskal–Wallis *H* test**[1] (named after William Kruskal and W. Allen Wallis), or **one-way ANOVA on ranks**[1] is a non-parametric method for testing whether samples originate from the same distribution.[2][3][4] It is used for comparing two or more independent samples of equal or different sample sizes. It extends the Mann–Whitney *U* test, which is used for comparing only two groups. The parametric equivalent of the Kruskal–Wallis test is the one-way analysis of variance (ANOVA).

A significant Kruskal–Wallis test indicates that at least one sample stochastically dominates one other sample. The test does not identify where this stochastic dominance occurs or for how many pairs of groups stochastic dominance obtains. For analyzing the specific sample pairs for stochastic dominance, Dunn's test,[5] pairwise Mann–Whitney tests with Bonferroni correction,[6] or the more powerful but less well known Conover–Iman test[6] are sometimes used.

Since it is a nonparametric method, the Kruskal–Wallis test does not assume a normal distribution of the residuals, unlike the analogous one-way analysis of variance. If the researcher can make the assumptions of an identically shaped and scaled distribution for all groups, except for any difference in medians, then the null hypothesis is that the medians of all groups are equal, and the alternative hypothesis is that at least one population median of one group is different from the population median of at least one other group. Otherwise, it is impossible to say, whether the rejection of the null hypothesis comes from the shift in locations or group dispersions. This is the same issue that happens also with the Mann-Whitney test.[7][8][9]

Contents [hide]

1 Method

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link

11 / 11

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:17
10.2 - Hypothesis Testing | ST | Kruskal–Wallis one-way analys | Anscombe's quartet - Wikiped | Bias of an estimator - Wikiped | EDA_FE.ipynb - Colaboratory | Markov's inequality - Wikiped | Chebyshev's inequality - Wiki
en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance
Update

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link

Kruskal–Wallis one-way analysis of variance

From Wikipedia, the free encyclopedia

The **Kruskal–Wallis test** by ranks, **Kruskal–Wallis *H* test**[1] (named after William Kruskal and W. Allen Wallis), or **one-way ANOVA on ranks**[1] is a non-parametric method for testing whether samples originate from the same distribution.[2][3][4] It is used for comparing two or more independent samples of equal or different sample sizes. It extends the Mann–Whitney *U* test, which is used for comparing only two groups. The parametric equivalent of the Kruskal–Wallis test is the one-way analysis of variance (ANOVA).

A significant Kruskal–Wallis test indicates that at least one sample stochastically dominates one other sample. The test does not identify where this stochastic dominance occurs or for how many pairs of groups stochastic dominance obtains. For analyzing the specific sample pairs for stochastic dominance, Dunn's test,[5] pairwise Mann–Whitney tests with Bonferroni correction,[6] or the more powerful but less well known Conover–Iman test[6] are sometimes used.

Since it is a nonparametric method, the Kruskal–Wallis test does not assume a normal distribution of the residuals, unlike the analogous one-way analysis of variance. If the researcher can make the assumptions of an identically shaped and scaled distribution for all groups, except for any difference in medians, then the null hypothesis is that the medians of all groups are equal, and the alternative hypothesis is that at least one population median of one group is different from the population median of at least one other group. Otherwise, it is impossible to say, whether the rejection of the null hypothesis comes from the shift in locations or group dispersions. This is the same issue that happens also with the Mann-Whitney test.[7][8][9]

Contents [hide]

12 / 12

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:18
10.2 - Hypothesis Testing | ST… Kruskal–Wallis one-way analys… Anscombe's quartet - Wikiped… Bias of an estimator - Wikiped… EDA_FE.ipynb - Colaboratory Markov's inequality - Wikiped… Chebyshev's inequality - Wiki…
en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance
Update

Download as PDF
Printable version

Languages
العربية
Deutsch
Español
Français
한국어
Italiano
Português
Русский
Türkçe
7 more
Edit links

Method [edit]

1. Rank all data from all groups together; i.e., rank the data from 1 to *N* ignoring group membership. Assign any tied values the average of the ranks they would have received had they not been tied.
2. The test statistic is given by:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}, \text{ where:}$$

- N is the total number of observations across all groups
- g is the number of groups
- n_i is the number of observations in group i
- r_{ij} is the rank (among all observations) of observation j from group i
- $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group i
- $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.

3. If the data contain no ties the denominator of the expression for H is exactly $(N-1)N(N+1)/12$ and $\bar{r} = \frac{N+1}{2}$. Thus

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\left(\bar{r}_{i\cdot} - \frac{N+1}{2}\right)^2$$

$$= \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\bar{r}_{i\cdot}^2 - 3(N+1)$$

The la…

N observations
g #gps

13 / 13

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:18
10.2 - Hypothesis Testing | ST... Kruskal–Wallis one-way analy... Anscombe's quartet - Wikiped... Bias of an estimator - Wikiped... EDA_FE.ipynb - Colaboratory Markov's inequality - Wikiped... Chebyshev's inequality - Wiki...
en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance
Update

Download as PDF
Printable version

Languages
العربية
Deutsch
Español
Français
한국어
Italiano
Português
Русский
Türkçе
7 more
Edit links

Method [edit]

1. Rank all data from all groups together; i.e., rank the data from 1 to N ignoring group membership. Assign any tied values the average of the ranks they would have received had they not been tied.
2. The test statistic is given by:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}, \text{ where:}$$

   - N is the total number of observations across all groups
   - g is the number of groups
   - n_i is the number of observations in group i
   - r_{ij} is the rank (among all observations) of observation j from group i
   - $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group i
   - $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.
3. If the data contain no ties the denominator of the expression for H is exactly $(N-1)N(N+1)/12$ and $\bar{r} = \frac{N+1}{2}$. Thus

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\left(\bar{r}_{i\cdot} - \frac{N+1}{2}\right)^2$$

$$= \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\bar{r}_{i\cdot}^2 - 3(N+1)$$

The la...

14 / 14

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:19

pearson corr. coeff x_i's & y_i's

vs

SRCC $\longrightarrow$ rank(x_i) & rank(y_i)

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:19
10.2 - Hypothesis Testing | ST... Kruskal–Wallis one-way analy... Anscombe's quartet - Wikiped... Bias of an estimator - Wikiped... EDA_FE.ipynb - Colaboratory Markov's inequality - Wikiped... Chebyshev's inequality - Wiki...
en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance
Update

Download as PDF
Printable version

Languages
العربية
Deutsch
Español
Français
한국어
Italiano
Português
Русский
Türkçe
7 more
Edit links

7 External links

Method [edit]

1. Rank all data from all groups together; i.e., rank the data from 1 to *N* ignoring group membership. Assign any tied values the average of the ranks they would have received had they not been tied.
2. The test statistic is given by:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}, \text{ where:}$$

- N is the total number of observations across all groups
- g is the number of groups
- n_i is the number of observations in group i
- r_{ij} is the rank (among all observations) of observation j from group i
- $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group i
- $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.

r_1
r_2
$\vdots$
r_N
across all groups

3. If the data contain no ties the denominator of the expression for H is exactly $(N-1)N(N+1)/12$ and $\bar{r} = \frac{N+1}{2}$. Thus

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\left(\bar{r}_{i\cdot} - \frac{N+1}{2}\right)^2$$

$$= \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\bar{r}_{i\cdot}^2 - 3(N+1)$$

The la...

16 / 16

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:20
10.2 - Hypothesis Testing | ST… | Kruskal–Wallis one-way analys… | Anscombe's quartet - Wikiped… | Bias of an estimator - Wikiped… | EDA_FE.ipynb - Colaboratory | Markov's inequality - Wikiped… | Chebyshev's inequality - Wiki…
en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance
Update

Download as PDF
Printable version

Languages
العربية
Deutsch
Español
Français
한국어
Italiano
Português
Русский
Türkçe
7 more
Edit links

7 External links

Method [edit]

1. Rank all data from all groups together; i.e., rank the data from 1 to *N* ignoring group membership. Assign any tied values the average of the ranks they would have received had they not been tied.
2. The test statistic is given by:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}, \text{ where:}$$

   - N is the total number of observations across all groups
   - g is the number of groups
   - n_i is the number of observations in group i
   - r_{ij} is the rank (among all observations) of observation j from group i
   - $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group i
   - $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.
3. If the data contain no ties the denominator of the expression for H is exactly $(N-1)N(N+1)/12$ and $\bar{r} = \frac{N+1}{2}$. Thus

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\left(\bar{r}_{i\cdot} - \frac{N+1}{2}\right)^2$$

$$= \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\bar{r}_{i\cdot}^2 - 3(N+1)$$

The la…

17 / 17

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:21
10.2 - Hypothesis Testing | ST… | Kruskal–Wallis one-way analys… | Anscombe's quartet - Wikiped… | Bias of an estimator - Wikiped… | EDA_FE.ipynb - Colaboratory | Markov's inequality - Wikiped… | Chebyshev's inequality - Wiki…
en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance
Update

Download as PDF
Printable version

Languages
العربية
Deutsch
Español
Français
한국어
Italiano
Português
Русский
Türkçe
7 more
Edit links

7 External links

Method [edit]

1. Rank all data from all groups together; i.e., rank the data from 1 to N ignoring group membership. Assign any tied values the average of the ranks they would have received had they not been tied.
2. The test statistic is given by:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}, \text{ where:}$$

f-stat

   - N is the total number of observations across all groups
   - g is the number of groups
   - n_i is the number of observations in group i
   - r_{ij} is the rank (among all observations) of observation j from group i
   - $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group i
   - $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.
3. If the data contain no ties the denominator of the expression for H is exactly $(N-1)N(N+1)/12$ and $\bar{r} = \frac{N+1}{2}$. Thus

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\left(\bar{r}_{i\cdot} - \frac{N+1}{2}\right)^2$$

$$= \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\bar{r}_{i\cdot}^2 - 3(N+1)$$

The la…

18 / 18

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:22
10.2 - Hypothesis Testing | ST… | Kruskal–Wallis one-way analy… | Anscombe's quartet - Wikiped… | Bias of an estimator - Wikiped… | EDA_FE.ipynb - Colaboratory | Markov's inequality - Wikiped… | Chebyshev's inequality - Wiki…
en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance
Update

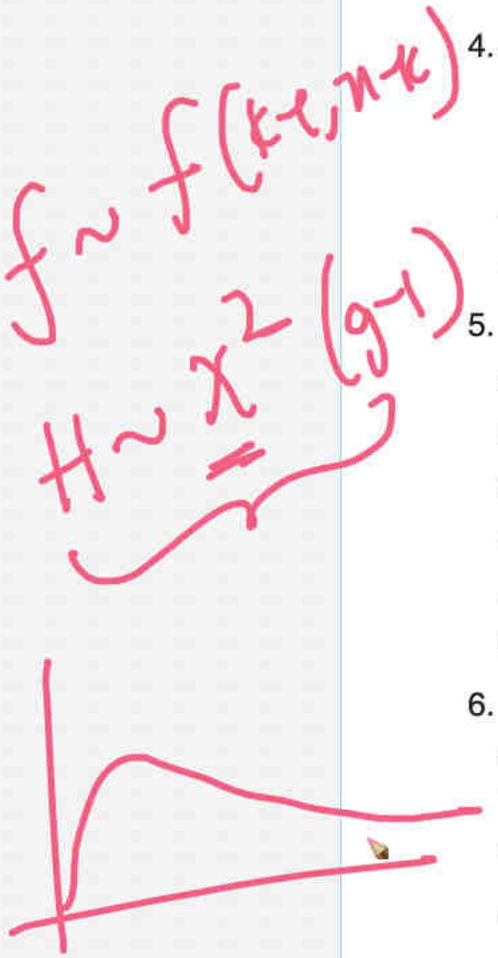
$$\frac{}{N(N+1)}\sum_{i=1}^{} n_i r_{i\cdot}^{} - 3(N+1)$$

The last formula only contains the squares of the average ranks.

4. A correction for ties if using the short-cut formula described in the previous point can be made by dividing H by $1 - \frac{\sum_{i=1}^{G}(t_i^3 - t_i)}{N^3 - N}$, where *G* is the number of groupings of different tied ranks, and t_i is the number of tied values within group *i* that are tied at a particular value. This correction usually makes little difference in the value of *H* unless there are a large number of ties.
5. Finally, the decision to reject or not the null hypothesis is made by comparing H to a critical value H_c obtained from a table or a software for a given significance or alpha level. If H is bigger than H_c, the null hypothesis is rejected. If possible (no ties, sample not too big) one should compare H to the critical value obtained from the exact distribution of H. Otherwise, the distribution of H can be approximated by a chi-squared distribution with g-1 degrees of freedom. If some n_i values are small (i.e., less than 5) the exact probability distribution of H can be quite different from this chi-squared distribution. If a table of the chi-squared probability distribution is available, the critical value of chi-squared, $\chi^2_{\alpha:g-1}$, can be found by entering the table at *g* – 1 degrees of freedom and looking under the desired significance or alpha level.
6. If the statistic is not significant, then there is no evidence of stochastic dominance between the samples. However, if the test is significant then at least one sample stochastically dominates another sample. Therefore, a researcher might use sample contrasts between individual sample pairs, or *post hoc* tests using Dunn's test, which (1) properly employs the same rankings as the Kruskal–Wallis test, and (2) properly employs the pooled variance implied by the null hypothesis of the Kruskal–Wallis test in order to determine which of the sample pairs are significantly different.[5] When performing multiple sample contrasts or tests, the Type I error rate tends to become inflated, raising concerns about multiple comparisons.

Exact probability tables [edit]

A large amount o… the Kruskal–Wallis test. Existing software only



Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:22
10.2 - Hypothesis Testing | ST | Kruskal–Wallis one-way analys | Anscombe's quartet - Wikiped | Bias of an estimator - Wikiped | EDA_FE.ipynb - Colaboratory | Markov's inequality - Wikiped | Chebyshev's inequality - Wiki
en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance
Update

Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Languages
العربية
Deutsch
Español
Français
한국어
Italiano
Português
Русский
Türkçe
7 more
Edit links

Method [edit]

1. Rank all data from all groups together; i.e., rank the data from 1 to *N* ignoring group membership. Assign any tied values the average of the ranks they would have received had they not been tied.
2. The test statistic is given by:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}, \text{ where:}$$

   - N is the total number of observations across all groups
   - g is the number of groups
   - n_i is the number of observations in group i
   - r_{ij} is the rank (among all observations) of observation j from group i
   - $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group i
   - $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.
3. If the data contain no ties the denominator of the expression for H is exactly $(N-1)N(N+1)/12$ and $\bar{r} = \frac{N+1}{2}$. Thus

$$H = \frac{12}{} \sum_{i=1}^{g} \left(\quad \frac{N+1}{} \right)^2$$

20 / 20

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:22
10.2 - Hypothesis Testing | ST | Kruskal–Wallis one-way analy | Anscombe's quartet - Wikiped | Bias of an estimator - Wikiped | EDA_FE.ipynb - Colaboratory | Markov's inequality - Wikiped | Chebyshev's inequality - Wiki
online.stat.psu.edu/stat200/lesson/10/10.2 Update

groups. A one-way between groups ANOVA comparing just two groups will give you the same results at the independent t test that you conducted in Lesson 8. We will use the five step hypothesis testing procedure again in this lesson.

1. Check assumptions and write hypotheses

The assumptions for a one-way between groups ANOVA are:

1. Samples are independent
2. The response variable is approximately normally distributed for each group or all group sample sizes are at least 30
3. The population variances are equal across responses for the group levels (if the largest sample standard deviation divided by the smallest sample standard deviation is **not** greater than **two**, then assume that the population variances are equal)

Given that you are comparing k independent groups, the null and alternative hypotheses are:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
H_a : Not all μ. are equal

In other words, the null hypothesis is that at all of the groups' population means are equal. The alternative is that they are not all equal; there are at least two population means that are not equal to one another.

2. Calculate the test statistic

ANOVA uses an F test statistic. Hand you will

Resources

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:23
10.2 - Hypothesis Testing | ST… Kruskal–Wallis one-way analy… Anscombe's quartet - Wikiped… Bias of an estimator - Wikiped… EDA_FE.ipynb - Colaboratory Markov's inequality - Wikiped… Chebyshev's inequality - Wiki…
en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance
Update

Printable version

Languages
العربية
Deutsch
Español
Français
한국어
Italiano
Português
Русский
Türkçe
7 more
Edit links

Method [edit]

1. Rank all data from all groups together; i.e., rank the data from 1 to *N* ignoring group membership. Assign any tied values the average of the ranks they would have received had they not been tied.
2. The test statistic is given by:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}, \text{ where:}$$

$\chi^2(g-1)$

- N is the total number of observations across all groups
- g is the number of groups
- n_i is the number of observations in group i
- r_{ij} is the rank (among all observations) of observation j from group i
- $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group i
- $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.

3. If the data contain no ties the denominator of the expression for H is exactly $(N-1)N(N+1)/12$ and $\bar{r} = \frac{N+1}{2}$. Thus

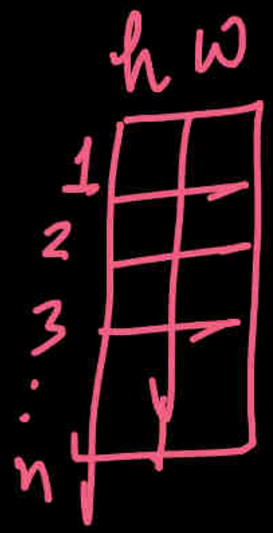
$$H = \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\left(\bar{r}_{i\cdot} - \frac{N+1}{2}\right)^2$$

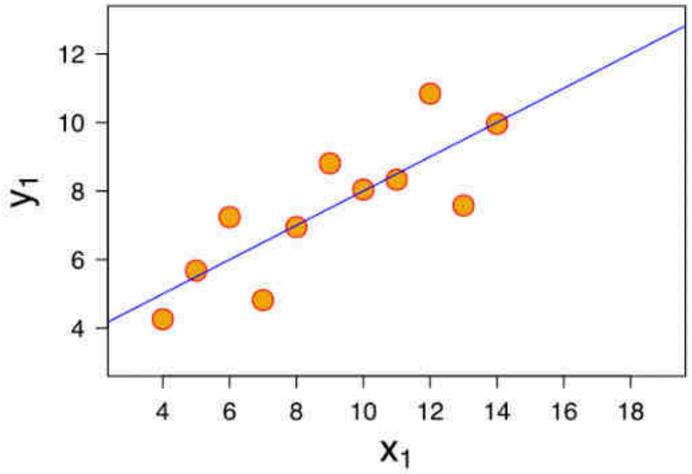
$$= \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\bar{r}_{i\cdot}^2 - 3(N+1)$$

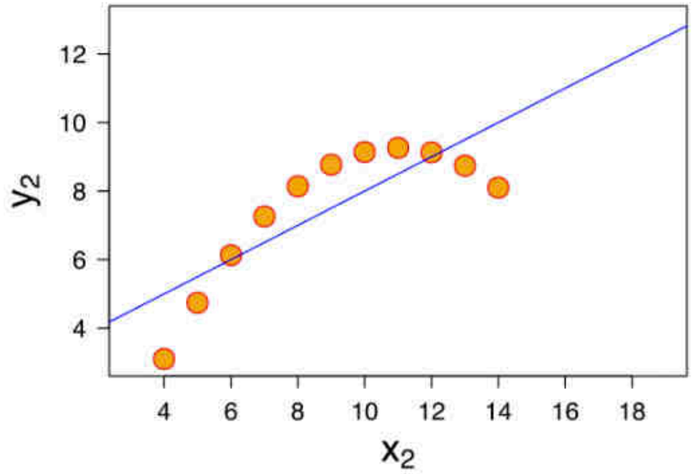
The la…

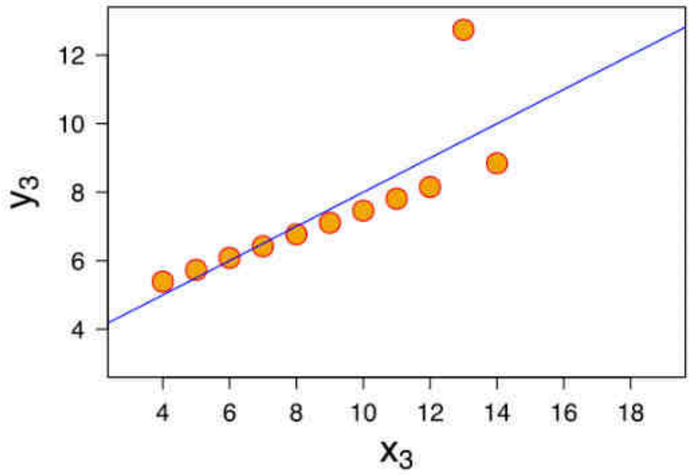
22 / 22

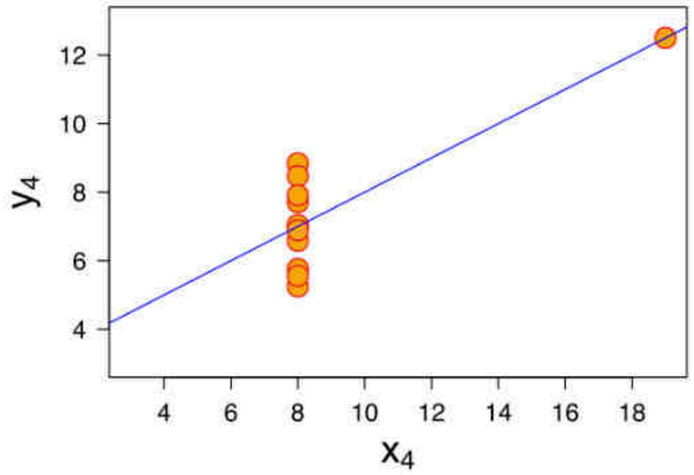
Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:28
10.2 - Hypothesis Testing | ST… | Kruskal–Wallis one-way analy… | Anscombe's quartet 3 - Ansc… | Bias of an estimator - Wikiped… | EDA_FE.ipynb - Colaboratory | Markov's inequality - Wikiped… | Chebyshev's inequality - Wiki…
en.wikipedia.org/wiki/Anscombe%27s_quartet#/media/File:Anscombe's_quartet_3.svg
Update











scatter plot

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

More details

Anscombe.svg: Schutz (label using subscri…

23 / 23

CC BY-SA 3.0

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:29
10.2 - Hypothesis Testing | ST… Kruskal–Wallis one-way analy… Anscombe's quartet - Wikiped… Bias of an estimator - Wikiped… EDA_FE.ipynb - Colaboratory Markov's inequality - Wikiped… Chebyshev's inequality - Wiki…
en.wikipedia.org/wiki/Anscombe%27s_quartet
Update

Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons

Languages
Deutsch
Español
Français
한국어
日本語
Português
Русский
Tiếng Việt
中文
11 more
Edit links

Data [edit]

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, ev…ween the variables.

24 / 24

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:29
10.2 - Hypothesis Testing | ST… | Kruskal–Wallis one-way analy… | Anscombe's quartet - Wikiped… | Bias of an estimator - Wikiped… | EDA_FE.ipynb - Colaboratory | Markov's inequality - Wikiped… | Chebyshev's inequality - Wiki…
en.wikipedia.org/wiki/Anscombe%27s_quartet
Update

Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons

Languages
Deutsch
Español
Français
한국어
日本語
Português
Русский
Tiếng Việt
中文
11 more
Edit links

Data [edit]

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

→ ML

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, ev…ween the variables.

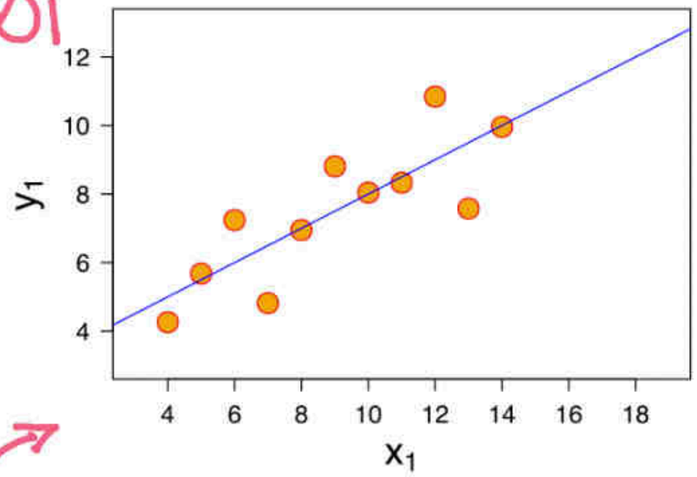
25 / 25

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:31
10.2 - Hypothesis Testing | ST… | Kruskal–Wallis one-way analy… | Anscombe's quartet 3 - Ansc… | Bias of an estimator - Wikiped… | EDA_FE.ipynb - Colaboratory | Markov's inequality - Wikiped… | Chebyshev's inequality - Wiki…
en.wikipedia.org/wiki/Anscombe%27s_quartet#/media/File:Anscombe's_quartet_3.svg
Update

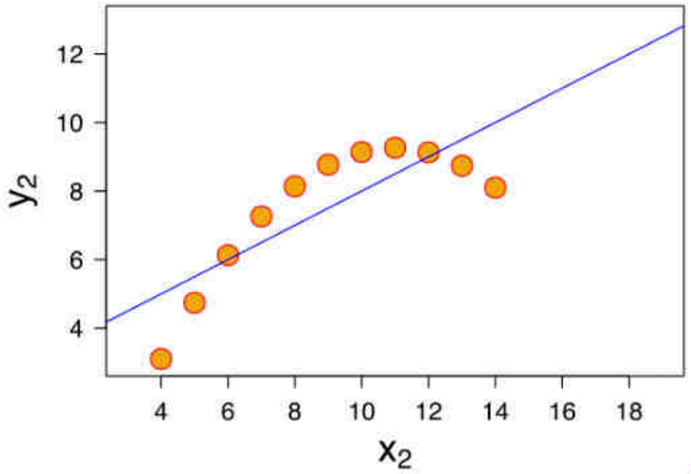
Stats

look at data

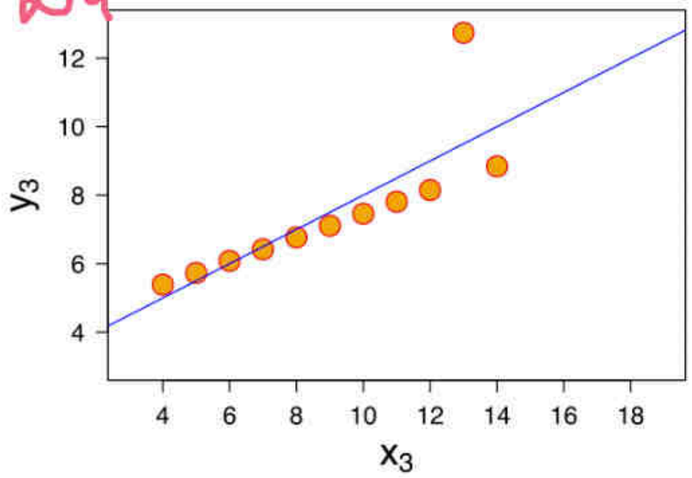
D1



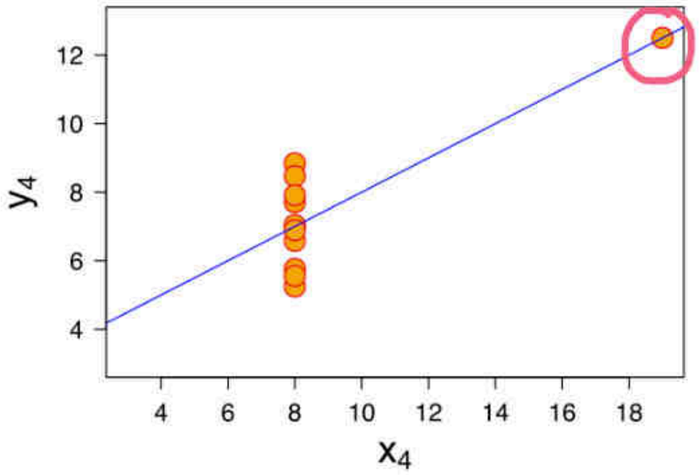
D2



D4



D3



0.9

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

More details

Anscombe.svg: Schutz (label using subscri…
26 / 26
CC BY-SA 3.0

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:39

Biased estimate

Data → param

$x_1 \dots x_n$

m

s

$\tilde{\mu}$

$\tilde{\sigma}$

$X : x_1 \cdots x_n$

estimate θ: param

param = θ

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:42
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | New Tab
en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance
Update

data follow some unknown distribution $P(x \mid \theta)$ (where θ is a fixed, unknown constant that is part of this distribution), and then we construct some estimator $\hat{\theta}$ that maps observed data to values that we hope are close to θ. The **bias** of $\hat{\theta}$ relative to θ is defined as[1]

$$\operatorname{Bias}(\hat{\theta}, \theta) = \operatorname{Bias}_\theta[\hat{\theta}] = \mathrm{E}_{x|\theta}[\hat{\theta}] - \theta = \mathrm{E}_{x|\theta}[\hat{\theta} - \theta],$$

where $\mathrm{E}_{x|\theta}$ denotes expected value over the distribution $P(x \mid \theta)$ (i.e., averaging over all possible observations x). The second equation follows since θ is measurable with respect to the conditional distribution $P(x \mid \theta)$.

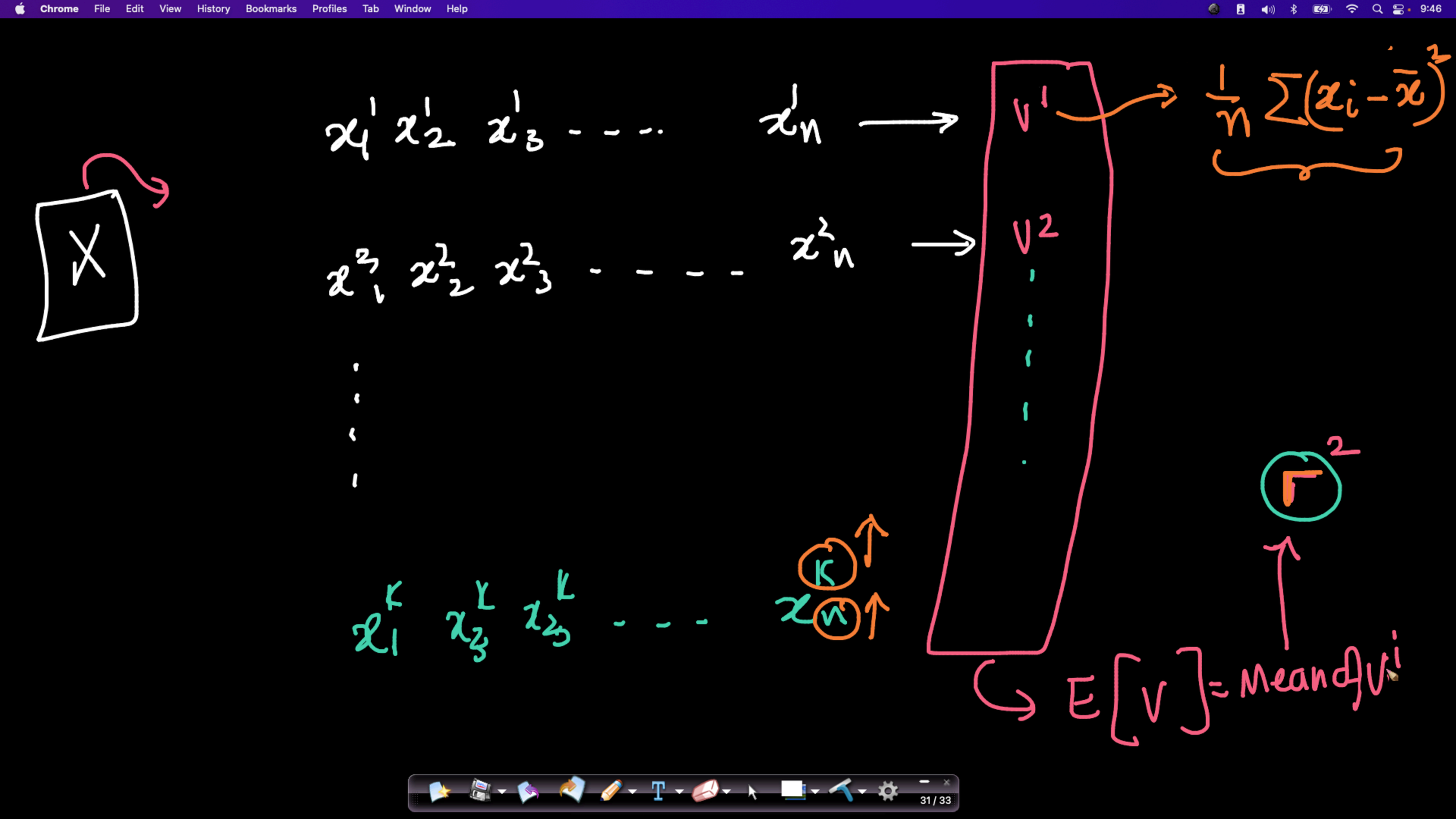
An estimator is said to be **unbiased** if its bias is equal to zero for all values of parameter θ, or equivalently, if the expected value of the estimator matches that of the parameter.[2]

In a simulation experiment concerning the properties of an estimator, the bias of the estimator may be assessed using the mean signed difference.

Examples [edit]

Sample variance [edit]

Ma


Chrome File Edit View History Bookmarks Profiles Tab Window Help
9:46
X
$x_1^1\ x_2^1\ x_3^1 \cdots x_n^1$
$x_1^2\ x_2^2\ x_3^2 \cdots x_n^2$
$x_1^K\ x_2^K\ x_3^K \cdots$
K
N
V1
V2
$\frac{1}{n}\sum(x_i - \bar{x})^2$
σ^2
E[V] = Mean of Vi
31 / 33

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:44

want:

$$E[s] = \sigma \quad \text{ideally}$$

$$\text{Bias}_{\sigma}(s) = \sigma - \bar{E}[s] \quad \text{as low as possible}$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:49

variance

if $var = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ on a sample of n-pts

this formula + Mallı

$E[s^2] \neq \sigma^2$

ideally $E[var] = \sigma^2$

But

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:47
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | New Tab
en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance
Update

Dividing instead by $n - 1$ yields an unbiased estimator. Conversely, MSE can be minimized by dividing by a different number (depending on distribution), but this results in a biased estimator. This number is always larger than $n - 1$, so this is known as a shrinkage estimator, as it "shrinks" the unbiased estimator towards zero; for the normal distribution the optimal value is $n + 1$.

Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables with expectation μ and variance σ^2. If the sample mean and uncorrected sample variance are defined as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

then S^2 is a biased estimator of σ^2, because

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left((X_i - \mu) - (\overline{X} - \mu)\right)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left((X_i - \mu)^2 - 2(\overline{X} - \mu)(X_i - \mu) + (\overline{X} - \mu)^2\right)\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n} (X_i - \mu) + \frac{1}{n}(\overline{X} - \mu)^2 \sum_{i=1}^{n} 1\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n} (X_i - \mu) + \frac{1}{n}(\overline{X} - \mu)^2 \cdot n\right]
\end{aligned}$$

34 / 34

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:47
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | New Tab
en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance
Update

Dividing instead by $n - 1$ yields an unbiased estimator. Conversely, MSE can be minimized by dividing by a different number (depending on distribution), but this results in a biased estimator. This number is always larger than $n - 1$, so this is known as a shrinkage estimator, as it "shrinks" the unbiased estimator towards zero; for the normal distribution the optimal value is $n + 1$.

Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables with expectation μ and variance σ^2. If the sample mean and uncorrected sample variance are defined as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

then S^2 is a biased estimator of σ^2, because

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left((X_i - \mu) - (\overline{X} - \mu)\right)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left((X_i - \mu)^2 - 2(\overline{X} - \mu)(X_i - \mu) + (\overline{X} - \mu)^2\right)\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n} (X_i - \mu) + \frac{1}{n}(\overline{X} - \mu)^2 \sum_{i=1}^{n} 1\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n} (X_i - \mu) + \frac{1}{n}(\overline{X} - \mu)^2 \cdot n\right]
\end{aligned}$$

35 / 35

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:48
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | New Tab
en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance
Update

then S^2 is a biased estimator of σ^2, because

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Big((X_i-\mu) - (\overline{X}-\mu)\Big)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left((X_i-\mu)^2 - 2(\overline{X}-\mu)(X_i-\mu) + (\overline{X}-\mu)^2\right)\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + \frac{1}{n}(\overline{X}-\mu)^2\sum_{i=1}^{n}1\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + \frac{1}{n}(\overline{X}-\mu)^2\cdot n\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + (\overline{X}-\mu)^2\right]
\end{aligned}$$

To continue, we note that by subtracting μ from both sides of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we get

$$\overline{X} - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu).$$

36 / 36

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:48
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | New Tab
en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance
Update

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - 2(\overline{X}-\mu)^2 + (\overline{X}-\mu)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\overline{X}-\mu)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] - \mathrm{E}\left[(\overline{X}-\mu)^2\right]$$

$$= \sigma^2 - \mathrm{E}\left[(\overline{X}-\mu)^2\right] = \left(1-\frac{1}{n}\right)\sigma^2 < \sigma^2.$$

$E[s^2] = \left(1-\frac{1}{n}\right)\sigma^2$

This can be seen by noting the following formula, which follows from the Bienaymé formula, for the term in the inequality for the expectation of the uncorrected sample variance above: $\mathrm{E}\left[(\overline{X}-\mu)^2\right] = \frac{1}{n}\sigma^2$.

In other words, the expected value of the uncorrected sample variance does not equal the population variance σ^2, unless multiplied by a normalization factor. The sample mean, on the other hand, is an unbiased[3] estimator of the population mean μ.[2]

Note that the usual definition of sample variance is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$, and this is an unbiased estimator of the population

37 / 37

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:49
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | New Tab
en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance
Update

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\overline{X}-\mu)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] - \mathrm{E}\left[(\overline{X}-\mu)^2\right]$$

$$= \sigma^2 - \mathrm{E}\left[(\overline{X}-\mu)^2\right] = \left(1-\frac{1}{n}\right)\sigma^2 < \sigma^2.$$

$E[S^2] = \sigma^2$

$\left(1-\frac{1}{n}\right)\sigma^2 + \sigma^2$

This can be seen by noting the following formula, which follows from the Bienaymé formula, for the term in the inequality for the expectation of the uncorrected sample variance above: $\mathrm{E}\left[(\overline{X}-\mu)^2\right] = \frac{1}{n}\sigma^2$.

In other words, the expected value of the uncorrected sample variance does not equal the population variance σ^2, unless multiplied by a normalization factor. The sample mean, on the other hand, is an unbiased[3] estimator of the population mean μ.[2]

Note that the usual definition of sample variance is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, and this is an unbiased estimator of the population variance.

Algebraically speaking, $\mathrm{E}[S^2]$ is unbiased because:

$$\mathrm{E}[S^2] = \left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right]$$

38 / 38

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:50
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | New Tab
en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance
Update

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] - \mathrm{E}\left[(\overline{X}-\mu)^2\right]$$

$$= \sigma^2 - \mathrm{E}\left[(\overline{X}-\mu)^2\right] = \left(1-\frac{1}{n}\right)\sigma^2 < \sigma^2.$$

This can be seen by noting the following formula, which follows from the Bienaymé formula, for the term in the inequality for the expectation of the uncorrected sample variance above: $\mathrm{E}\left[(\overline{X}-\mu)^2\right] = \frac{1}{n}\sigma^2$.

In other words, the expected value of the uncorrected sample variance does not equal the population variance σ^2, unless multiplied by a normalization factor. The sample mean, on the other hand, is an unbiased[3] estimator of the population mean μ.[2]

Note that the usual definition of sample variance is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$, and this is an unbiased estimator of the population variance.

Algebraically speaking, $\mathrm{E}[S^2]$ is unbiased because:

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right] = \frac{n}{n-1}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right]$$

$$=$$

39 / 39

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:50
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | New Tab
en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance Update

for the expectation of the uncorrected sample variance above: $\mathrm{E}\left[(\overline{X}-\mu)^2\right] = \frac{1}{n}\sigma^2$.

In other words, the expected value of the uncorrected sample variance does not equal the population variance σ^2, unless multiplied by a normalization factor. The sample mean, on the other hand, is an unbiased[3] estimator of the population mean μ.[2]

Note that the usual definition of sample variance is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, and this is an unbiased estimator of the population variance.

Algebraically speaking, $\mathrm{E}[S^2]$ is unbiased because:

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] = \frac{n}{n-1}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]$$

$$= \frac{n}{n-1}\left(1 - \frac{1}{n}\right)\sigma^2 = \sigma^2,$$

$n \to \infty$

$\mathrm{E}[S^2] = \sigma^2$

where the transition to the second line uses the result derived above for the biased estimator. Thus $\mathrm{E}[S^2] = \sigma^2$, and therefore $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimator of the population variance, σ^2. The ratio between the biased (uncorrected) and unbiased estimates of the variance is known as Bessel's correction.

The reason tha[illegible] the fact that the sample mean is an ordinary

40 / 40

Var:

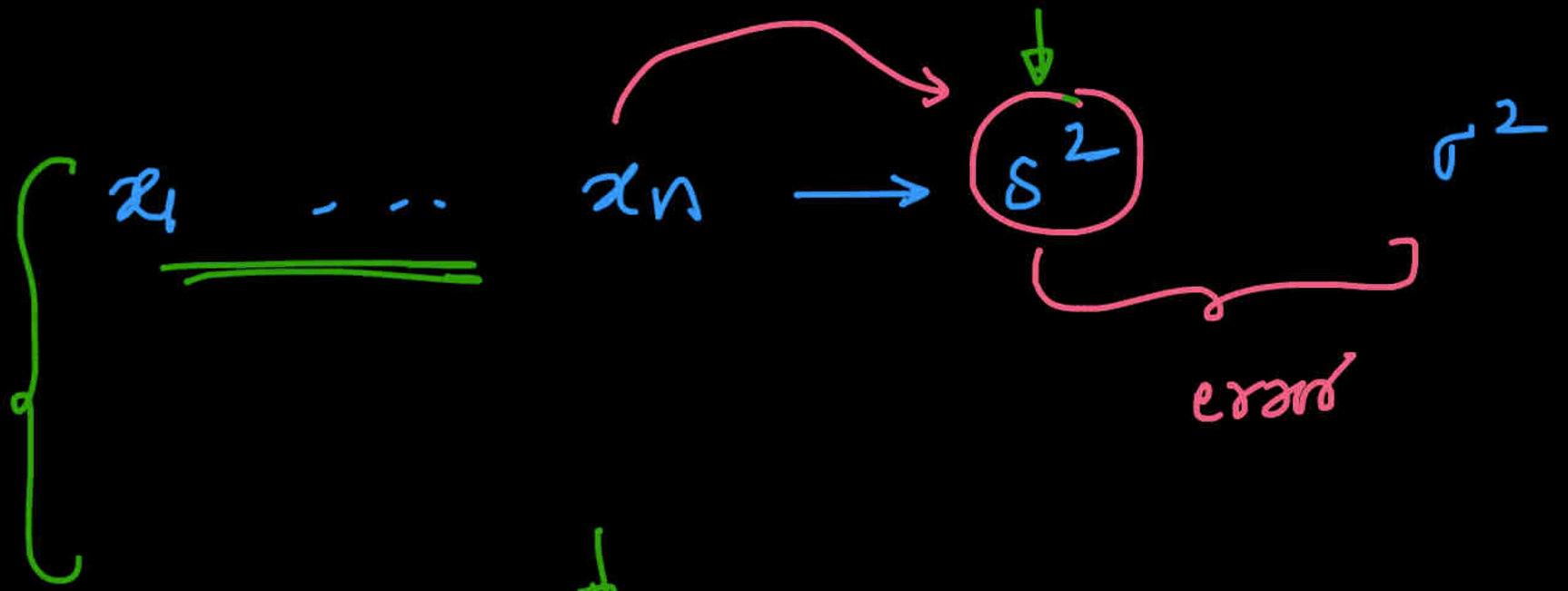
$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

if n is small $\Rightarrow$ n & $n-1$ will change the value

as $n\uparrow$ diff $n-1$ & n doesn't matter as much

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:53

Bias - variance tradeoff → ML module



Bias! → $E[s^2] - \sigma^2$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:59

10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=A-B0OnWTn1WV

Update

+ Code + Text Reconnect

Task: Determine the eligibility for granting Home loan.

Objective of this notebook is:

1. To understand the patterns in the data.
2. How to Handle the categorical features.
3. How to deal with missing data.
4. Feature Engineering
5. Finding the most important features while taking the decision of granting a loan application.
6. Understanding the Normalization and standardisation of the data.

Load data and libraries

[ ] ↳ 11 cells hidden

Basic Data Exploration

[ ] ↳ 9 cells hidden

44 / 44

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:00

10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=A-B0OnWTn1WV Update

+ Code + Text Reconnect

Task: Determine the eligibility for granting Home loan.

Objective of this notebook is:

1. To understand the patterns in the data.
2. How to Handle the categorical features.
3. How to deal with missing data.
4. Feature Engineering
5. Finding the most important features while taking the decision of granting a loan application.
6. Understanding the Normalization and standardisation of the data.

Load data and libraries

[ ] ↳ 11 cells hidden

Basic Data Exploration

[ ] ↳ 9 cells hidden

45 / 45

```
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')
```

```
#Data: https://drive.google.com/file/d/1oJbdRpTLqPu1SIBXHkzWRaLaZbvZot7w/view?usp=sharing
# Download data
id = "1oJbdRpTLqPu1SIBXHkzWRaLaZbvZot7w"
path = "https://docs.google.com/uc?export=download&id=" + id
print(path)
```

```
https://docs.google.com/uc?export=download&id=1oJbdRpTLqPu1SIBXHkzWRaLaZbvZot7w
```

```
!wget "https://docs.google.com/uc?export=download&id=1oJbdRpTLqPu1SIBXHkzWRaLaZbvZot7w" -O train.csv
```

```
--2022-07-25 13:37:17--  https://docs.google.com/uc?export=download&id=1oJbdRpTLqPu1SIBXHkzWRaLaZbvZot7w
Resolving docs.google.com (docs.google.com)... 172.253.62.138, 172.253.62.102, 172.253.62.100, ...
Connecting to docs.google.com (docs.google.com)|172.253.62.138|:443... connected.
HTTP request sent, awaiting response... 303 See Other
Location: https://doc-0o-90-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/d2lf91u0ibh30d
Warning: wildcards not supported in HTTP.
--2022-07-25 13:37:17--  https://doc-0o-90-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1
Resolving doc-0o-90-docs.googleusercontent.com (doc-0o-90-docs.googleusercontent.com)... 172.217.15.65, 2607:f8b0:40
Connecting to doc-0o-90-docs.googleusercontent.com (doc-0o-90-docs.googleusercontent.com)|172.217.15.65|:443... conn
HTTP request sent, awaiting response... 200 OK
Length: 38011 (37K) [text/csv]
```

```
data = pd.read_csv('./train.csv')
data.shape
```

```
(614, 13)
```

```
data.columns
```

```
Index(['Loan_ID', 'Gender', 'Married', 'Dependents', 'Education',
       'Self_Employed', 'ApplicantIncome', 'CoapplicantIncome', 'LoanAmount',
       'Loan_Amount_Term', 'Credit_History', 'Property_Area', 'Loan_Status'],
      dtype='object')
```

```
data.head()
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loa
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:02

10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=Ej8pWNLIGL-i

Update

+ Code + Text Reconnect

```
data = pd.read_csv('./train.csv')
data.shape
```

```
(614, 13)
```

```
[ ] data.columns
```

```
Index(['Loan_ID', 'Gender', 'Married', 'Dependents', 'Education',
       'Self_Employed', 'ApplicantIncome', 'CoapplicantIncome', 'LoanAmount',
       'Loan_Amount_Term', 'Credit_History', 'Property_Area', 'Loan_Status'],
      dtype='object')
```

```
data.head()
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loa
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	

48 / 48

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:03
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=Ej8pWNLIGL-i
Update
+ Code + Text Reconnect

```
[ ] data.dtypes
    #object => typically categorical/IDs
    #Int64, Float64
```

```
Loan_ID               object
Gender                object
Married               object
Dependents            object
Education             object
Self_Employed         object
ApplicantIncome        int64
CoapplicantIncome    float64
LoanAmount           float64
Loan_Amount_Term     float64
Credit_History       float64
Property_Area         object
Loan_Status           object
dtype: object
```

feature → categorical (discrete r.v)

```
[ ] data['Dependents'].value_counts()
```

```
0     345
1     102
2     101
```

49 / 49

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:04
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=Ej8pWNLIGL-i
Update
+ Code + Text Reconnect

```
[ ]  data.dtypes
     #object => typically categorical/IDs
     #Int64, Float64
```

```
Loan_ID              object
Gender               object
Married              object
Dependents           object
Education            object
Self_Employed        object
ApplicantIncome       int64
CoapplicantIncome   float64
LoanAmount          float64
Loan_Amount_Term    float64
Credit_History      float64
Property_Area        object
Loan_Status          object
dtype: object
```

```
[ ]  data['Dependents'].value_counts()
```

```
0     345
1     102
2     101
```

50 / 50

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:04
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=Ej8pWNLIGL-i
Update
+ Code + Text
Reconnect

```
[ ] data.dtypes
    #object => typically categorical/IDs
    #Int64, Float64
```

```
Loan_ID               object
Gender                object
Married               object
Dependents            object
Education             object
Self_Employed         object
ApplicantIncome        int64
CoapplicantIncome    float64
LoanAmount           float64
Loan_Amount_Term     float64
Credit_History       float64
Property_Area         object
Loan_Status           object
dtype: object
```

```
[ ] data['Dependents'].value_counts()
```

```
0     345
1     102
2     101
```

51 / 51

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:06


```
dtype: object
```

```
[ ] data['Dependents'].value_counts()
```

```
0     345
1     102
2     101
3+     51
Name: Dependents, dtype: int64
```


```
[ ] # drop loanID column
    data = data.drop('Loan_ID',axis = 1)
```

Basic Data Exploration

[ ] ↳ 9 cells hidden

Basic Data visualization: Univariate

[ ] ↳ 11 cells hidden

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:07

10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=f5n7flegGQEe

Update

+ Code + Text Reconnect

5. Understanding the Normalization and standardisation of the data.

Load data and libraries

↳ 11 cells hidden

Basic Data Exploration

[ ] ↳ 9 cells hidden

SRCC

dep. numeric vs income numeric

Basic Data visualization: Univariate

[ ] ↳ 11 cells hidden

Simple Feature Engineering

[ ] ↳ 37 cells hidden

53 / 54

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:09
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=TarGSsXDG8Tp
Update
+ Code + Text
Reconnect

Basic Data Exploration

```
[ ] data.describe()
    # only numeric features
```

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

```
[ ] # categorical features
```

54 / 55

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:09
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=TarGSsXDG8Tp
Update
+ Code + Text
Reconnect

Basic Data Exploration

```
[ ]  data.describe()
     # only numeric features
```

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

```
[ ]  # categorical features
```

55 / 56

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:09
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=TarGSsXDG8Tp
Update
+ Code + Text
Reconnect

Basic Data Exploration

```
[ ] data.describe()
    # only numeric features
```

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

```
[ ] # categorical features
```

56 / 57

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:10

10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=TarGSsXDG8Tp

Update

+ Code + Text Reconnect

Basic Data Exploration

```
[ ] data.describe()
    # only numeric features
```

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

```
[ ] # categorical features
```

57 / 58

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:12

max	81000.000000	41667.000000	700.000000	480.00000	1.000000	

```
[ ] # catgeorical features
    data.describe(include = ['object'])
```

	Gender	Married	Dependents	Education	Self_Employed	Property_Area	Loan_Status
count	601	611	599	614	582	614	614
unique	2	2	4	2	2	3	2
top	Male	Yes	0	Graduate	No	Semiurban	Y
freq	489	398	345	480	500	233	422

```
[ ] #missing values
    data.isna().sum()
```

```
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:12

max	81000.000000	41667.000000	700.000000	480.00000	1.000000

```
[ ] # catgeorical features
    data.describe(include = ['object'])
```

	Gender	Married	Dependents	Education	Self_Employed	Property_Area	Loan_Status
count	601	611	599	614	582	614	614
unique	2	2	4	2	2	3	2
top	Male	Yes	0	Graduate	No	Semiurban	Y
freq	489	398	345	480	500	233	422

```
[ ] #missing values
    data.isna().sum()
```

```
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:12

max	81000.000000	41667.000000	700.000000	480.00000	1.000000

```
[ ] # catgeorical features
    data.describe(include = ['object'])
```

	Gender	Married	Dependents	Education	Self_Employed	Property_Area	Loan_Status
count	601	611	599	614	582	614	614
unique	2	2	4	2	2	3	2
top	Male	Yes	0	Graduate	No	Semiurban	Y
freq	489	398	345	480	500	233	422

```
[ ] #missing values
    data.isna().sum()
```

```
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:12

max	81000.000000	41667.000000	700.000000	480.00000	1.000000

```
[ ] # catgeorical features
    data.describe(include = ['object'])
```

	Gender	Married	Dependents	Education	Self_Employed	Property_Area	Loan_Status
count	601	611	599	614	582	614	614
unique	2	2	4	2	2	3	2
top	Male	Yes	0	Graduate	No	Semiurban	Y
freq	489	398	345	480	500	233	422

```
[ ] #missing values
    data.isna().sum()
```

```
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:13

	Gender	Married	Dependents	Education	Self_Employed	Property_Area	Loan_Status
count	601	611	599	614	582	614	614
unique	2	2	4	2	2	3	2
top	Male	Yes	0	Graduate	No	Semiurban	Y
freq	489	398	345	480	500	233	422

```
#missing values
data.isna().sum()
```

```
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
Loan_Amount_Term    14
Credit_History      50
Property_Area        0
Loan_Status          0
dtype: int64
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:13
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=1YAwMv3fHa10 Update
+ Code + Text Reconnect

	Gender	Married	Dependents	Education	Self_Employed	Property_Area	Loan_Status
count	601	611	599	614	582	614	614
unique	2	2	4	2	2	3	2
top	Male	Yes	0	Graduate	No	Semiurban	Y
freq	489	398	345	480	500	233	422

```
#missing values
data.isna().sum()
```

```
Gender               13
Married               3
Dependents           15
Education             0
Self_Employed        32
ApplicantIncome       0
CoapplicantIncome     0
LoanAmount           22
Loan_Amount_Term     14
Credit_History       50
Property_Area         0
Loan_Status           0
dtype: int64
```

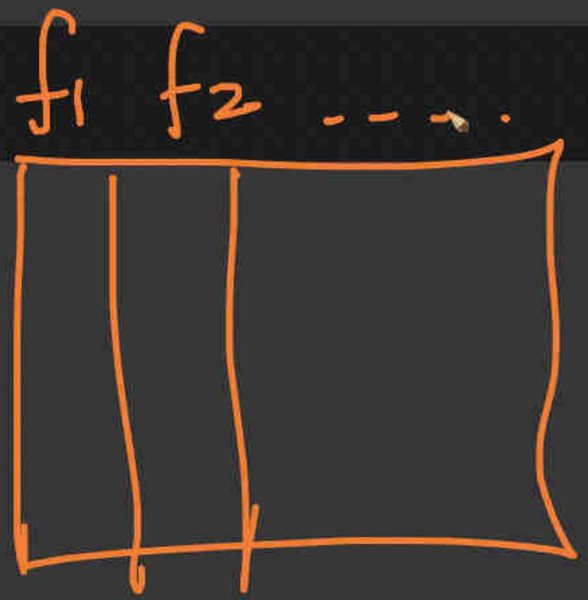
63 / 64

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:14
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=1YAwMv3fHa10 Update
+ Code + Text Reconnect

	Gender	Married	Dependents	Education	Self_Employed	Property_Area	Loan_Status
count	601	611	599	614	582	614	614
unique	2	2	4	2	2	3	2
top	Male	Yes	0	Graduate	No	Semiurban	Y
freq	489	398	345	480	500	233	422

```
#missing values
data.isna().sum()
```

```
Gender               13
Married               3
Dependents           15
Education             0
Self_Employed        32
ApplicantIncome       0
CoapplicantIncome     0
LoanAmount           22
Loan_Amount_Term     14
Credit_History       50
Property_Area         0
Loan_Status           0
dtype: int64
```



64 / 65

Chrome File Edit View History Bookmarks Profiles Tab Window Help
10:15
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=1YAwMv3fHa10
Update
+ Code + Text
Reconnect

	Gender	Married	Dependents	Education	Self_Employed	Property_Area	Loan_Status
count	601	611	599	614	582	614	614
unique	2	2	4	2	2	3	2
top	Male	Yes	0	Graduate	No	Semiurban	Y
freq	489	398	345	480	500	233	422

```
#missing values
data.isna().sum()
```

```
Gender               13
Married               3
Dependents           15
Education             0
Self_Employed        32
ApplicantIncome       0
CoapplicantIncome     0
LoanAmount           22
Loan_Amount_Term     14
Credit_History       50
Property_Area         0
Loan_Status           0
dtype: int64
```

65 / 66

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:16
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=1YAwMv3fHa10 Update
+ Code + Text Reconnect

unique	2	2	4	2	2	3	2
top	Male	Yes	0	Graduate	No	Semiurban	Y
freq	489	398	345	480	500	233	422

```
#missing values
data.isna().sum()
```

```
Gender               13
Married               3
Dependents           15
Education             0
Self_Employed        32
ApplicantIncome       0
CoapplicantIncome     0
LoanAmount           22
Loan_Amount_Term     14
Credit_History       50
Property_Area         0
Loan_Status           0
dtype: int64
```


```
[ ] # catgeorical and numerical columns
    cat_cols = data.dtypes =='object'
    cat_cols = list(cat_cols[cat_cols].index)
```

66 / 67

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:18

10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=1YAwMv3fHa10 Update

+ Code + Text Reconnect

```
dtype: int64
```

```
[ ] # catgeorical and numerical columns
    cat_cols = data.dtypes =='object'
    cat_cols = list(cat_cols[cat_cols].index)

    num_cols = data.dtypes !='object'
    num_cols = list(num_cols[num_cols].index)
    cat_cols.remove('Loan_Status')
```

```
[ ] cat_cols
```

```
['Gender',
 'Married',
 'Dependents',
 'Education',
 'Self_Employed',
 'Property_Area']
```

```
[ ] num_cols
```

```
['ApplicantIncome',
 'CoapplicantIncome',
 'LoanAmount',
 'Loan_Amount_Term',
```

67 / 68

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:18


```
dtype: int64
```

```
[ ] # catgeorical and numerical columns
    cat_cols = data.dtypes =='object'
    cat_cols = list(cat_cols[cat_cols].index)

    num_cols = data.dtypes !='object'
    num_cols = list(num_cols[num_cols].index)
    cat_cols.remove('Loan_Status')
```

```
[ ] cat_cols
```

```
['Gender',
 'Married',
 'Dependents',
 'Education',
 'Self_Employed',
 'Property_Area']
```

```
[ ] num_cols
```

```
['ApplicantIncome',
 'CoapplicantIncome',
 'LoanAmount',
 'Loan_Amount_Term',
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:19
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=GNmtX3um9oJR
Update

+ Code + Text Reconnect

Basic Data Exploration

[ ] ↳ 9 cells hidden

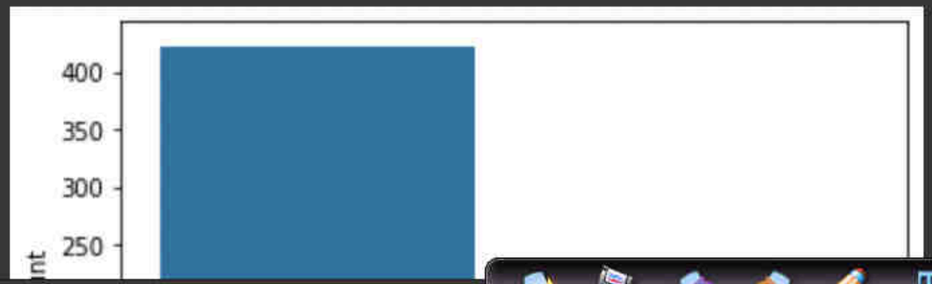
Basic Data visualization: Univariate

imbalanced dataset

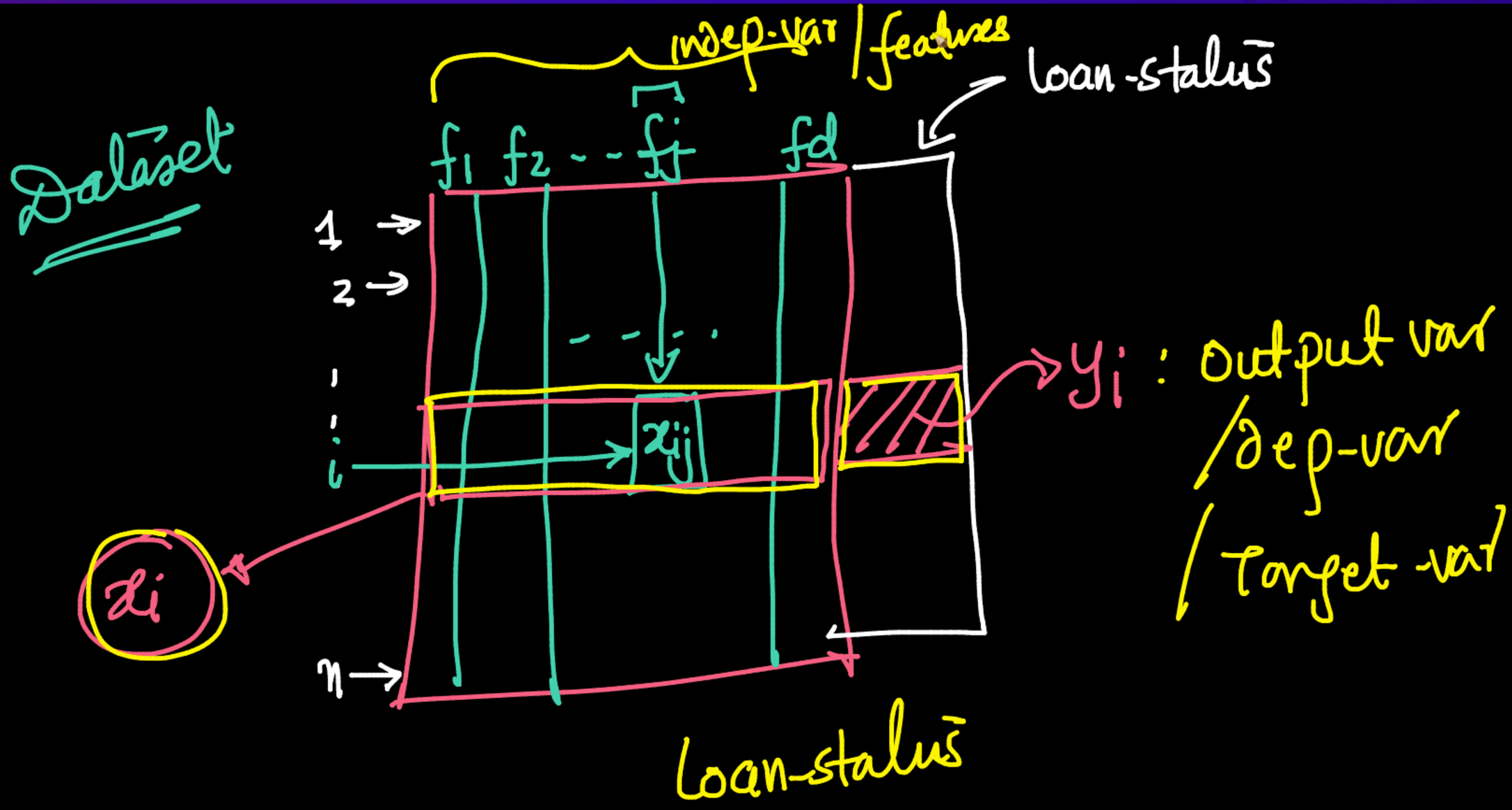
```
data['Loan_Status'].value_counts()
```

```
Y    422
N    192
Name: Loan_Status, dtype: int64
```

```
#Q: How many loans the company has approved in the past?
sns.countplot(data=data, x='Loan_Status')
plt.show()
```



Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:22



70 / 70


$y_i \in \{Yes, No\}$
binary r.v

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:23

10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Cola | Markov's inequality | Chebyshev's inequali | scipy.stats.kruskal | numpy.var — NumPy

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=GNmtX3um9oJR

Update

+ Code + Text Reconnect

Basic Data Exploration

[ ] ↳ 9 cells hidden

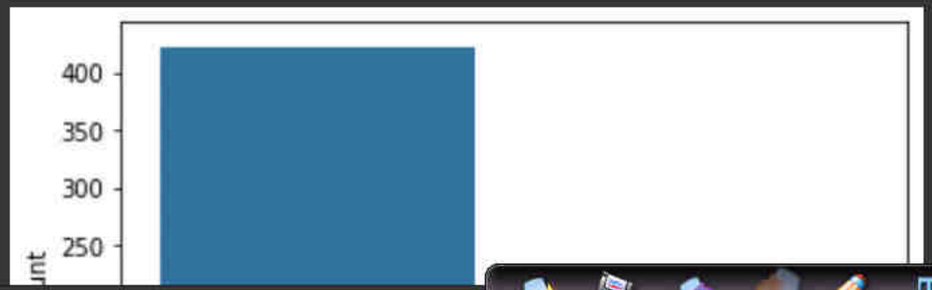
Basic Data visualization: Univariate

```
data['Loan_Status'].value_counts()
```

```
Y    422
N    192
Name: Loan_Status, dtype: int64
```

imbalanced dataset

```
[ ] #Q: How many loans the company has approved in the past?
    sns.countplot(data=data, x='Loan_Status')
    plt.show()
```



72 / 72

Chrome File Edit View History Bookmarks Profiles Tab Window Help

+ Code + Text

Reconnect

Basic Data Exploration

1 : 1000

[ ] ↳ 9 cells hidden

Basic Data visualization: Univariate

1 Questions

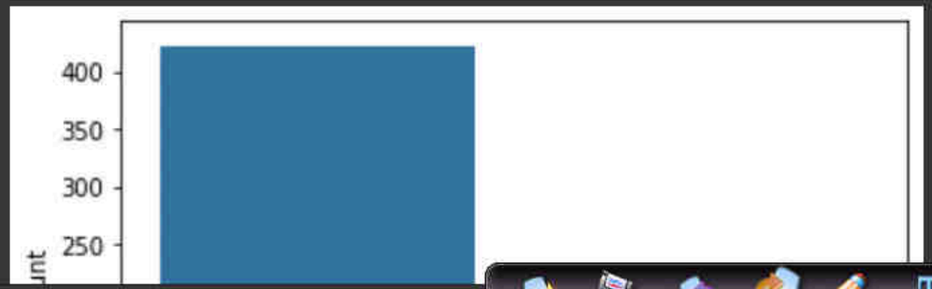
2 Techniques

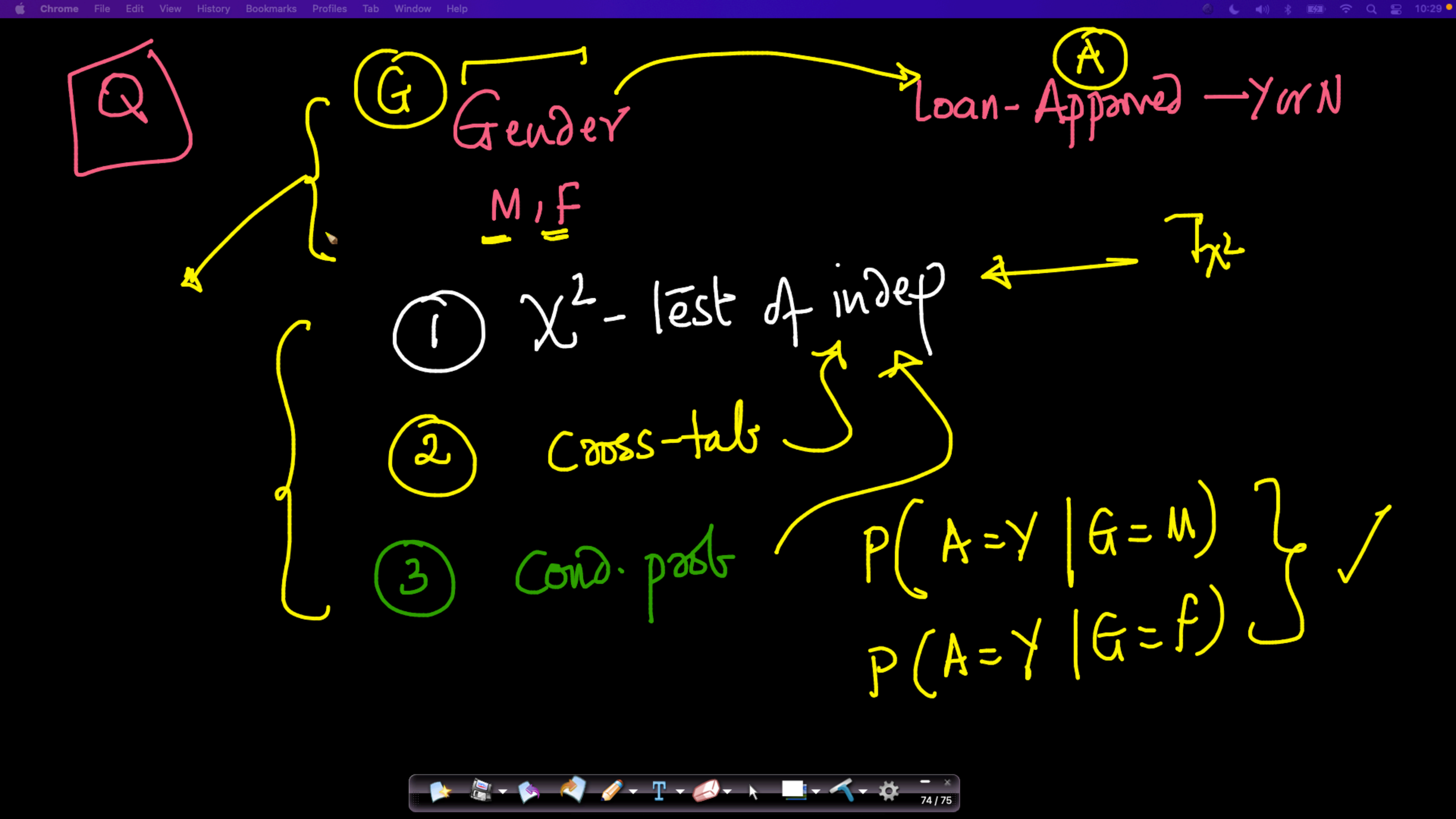
3 Answers

```
data['Loan_Status'].value_counts()
```

```
Y    422
N    192
Name: Loan_Status, dtype: int64
```

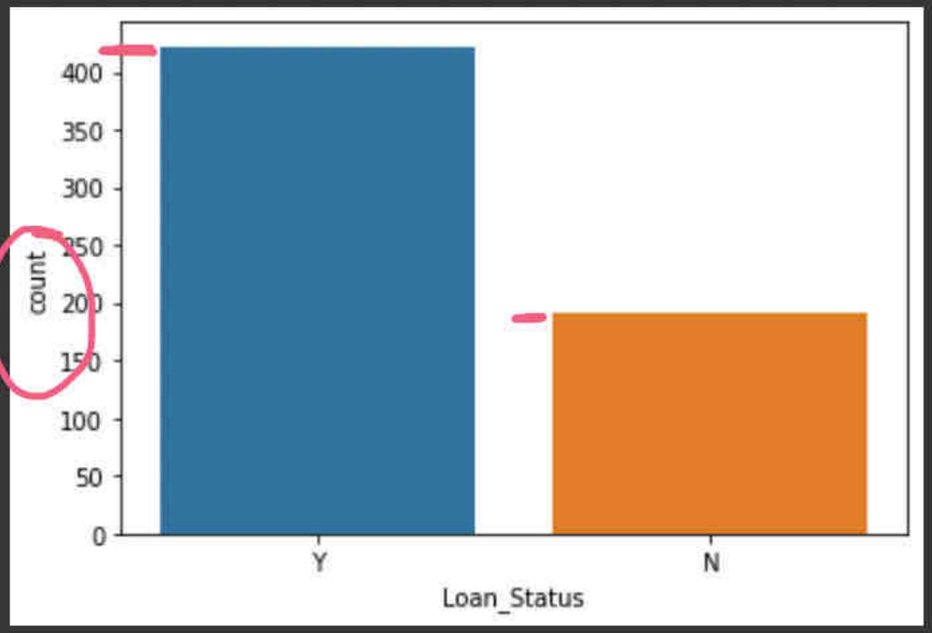
```
#Q: How many loans the company has approved in the past?
sns.countplot(data=data, x='Loan_Status')
plt.show()
```




Q
G
Gender
M, F
A
Loan - Approved → Y or N
1. χ^2 - test of indep
2. Cross-tab
3. Cond. prob
χ^2
$P(A=Y \mid G=M)$
$P(A=Y \mid G=F)$
✓

```
Y    422
N    192
Name: Loan_Status, dtype: int64
```

```
#Q: How many loans the company has approved in the past?
sns.countplot(data=data, x='Loan_Status')
plt.show()
```

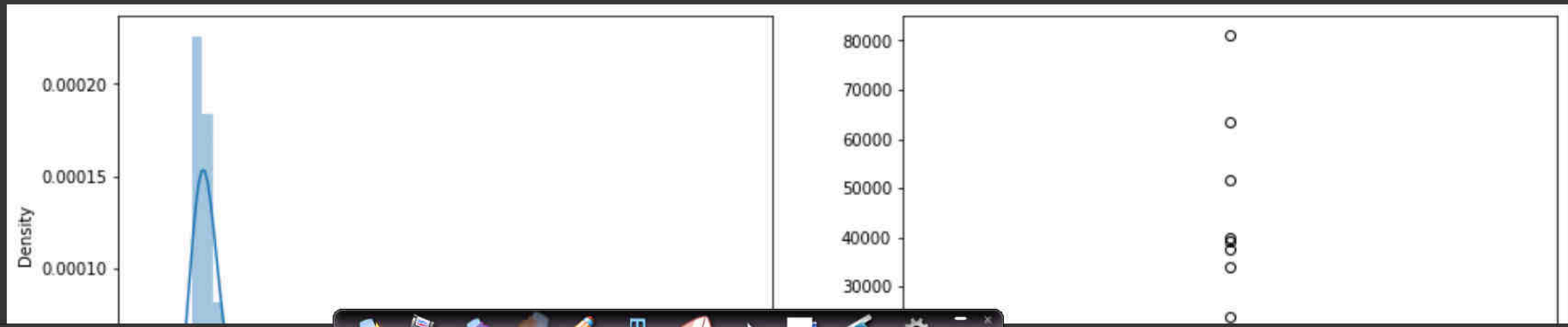



```
[ ]  target = 'Loan_Status'
     data[target].value_counts()
```

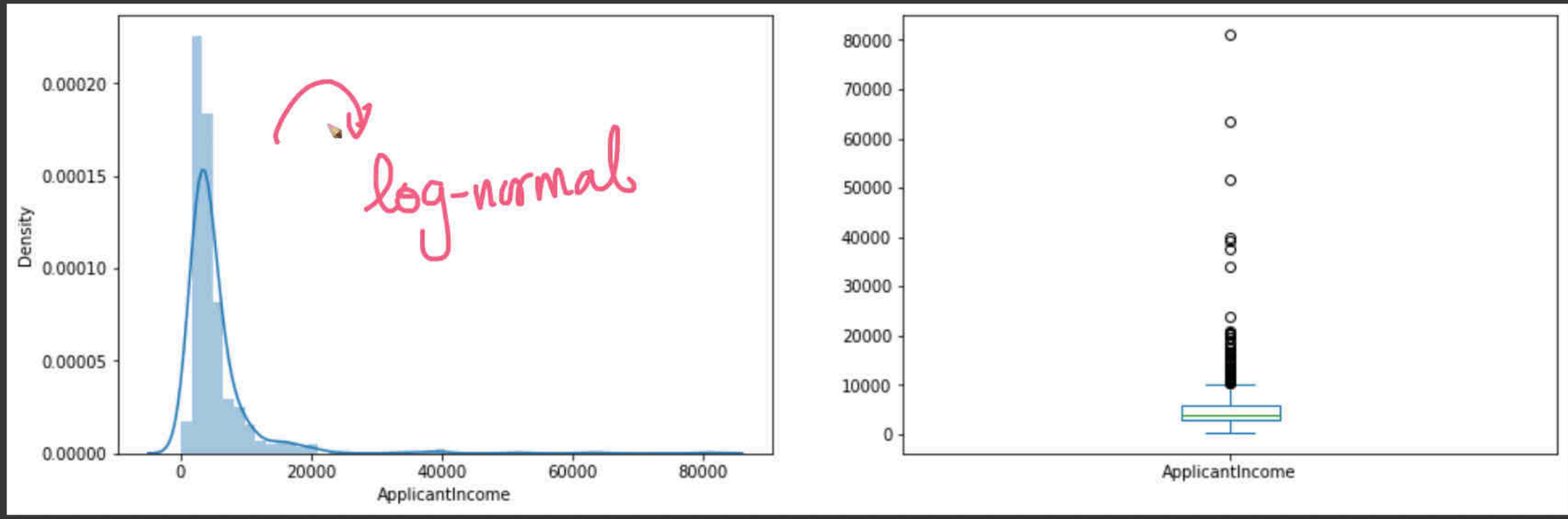
```
# Imbalanced data
```

```
Y    422
N    192
Name: Loan_Status, dtype: int64
```

```
#Income of the applicant
plt.subplot(121)
sns.distplot(data["ApplicantIncome"])

plt.subplot(122)
data["ApplicantIncome"].plot.box(figsize=(16,5))
plt.show()
```

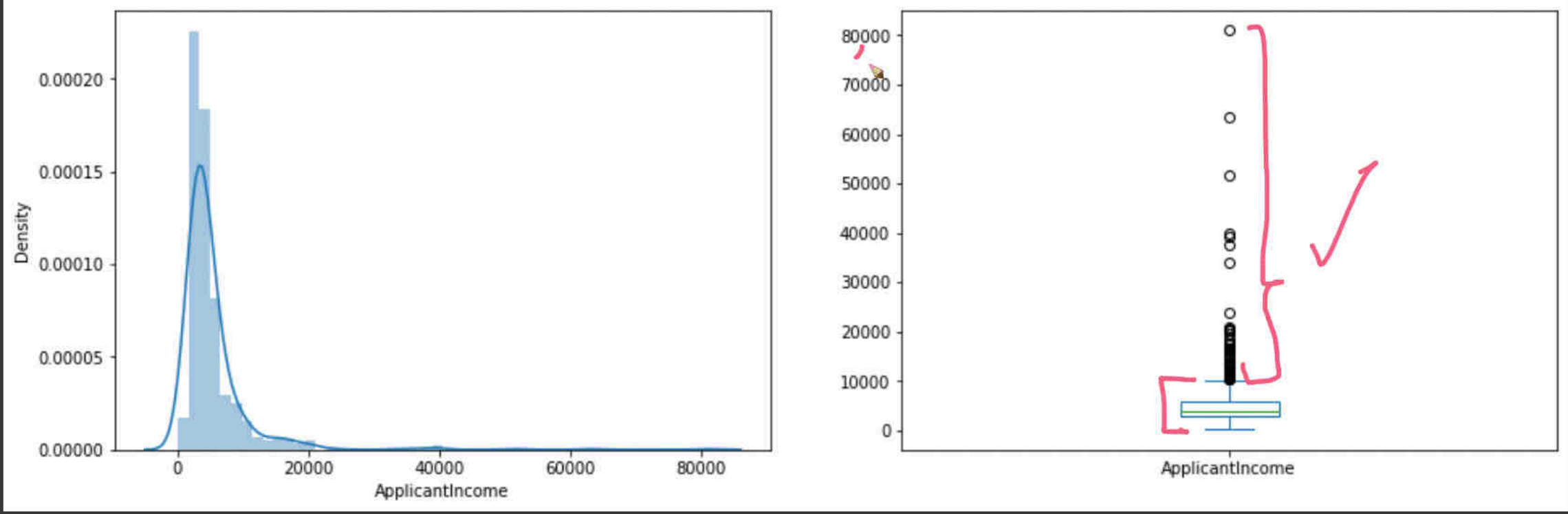


```
plt.subplot(122)
data["ApplicantIncome"].plot.box(figsize=(16,5))
plt.show()
```




```
[ ] plt.subplot(121)
    sns.distplot(np.log(data["ApplicantIncome"]))
```

```
plt.subplot(122)
data["ApplicantIncome"].plot.box(figsize=(16,5))
plt.show()
```

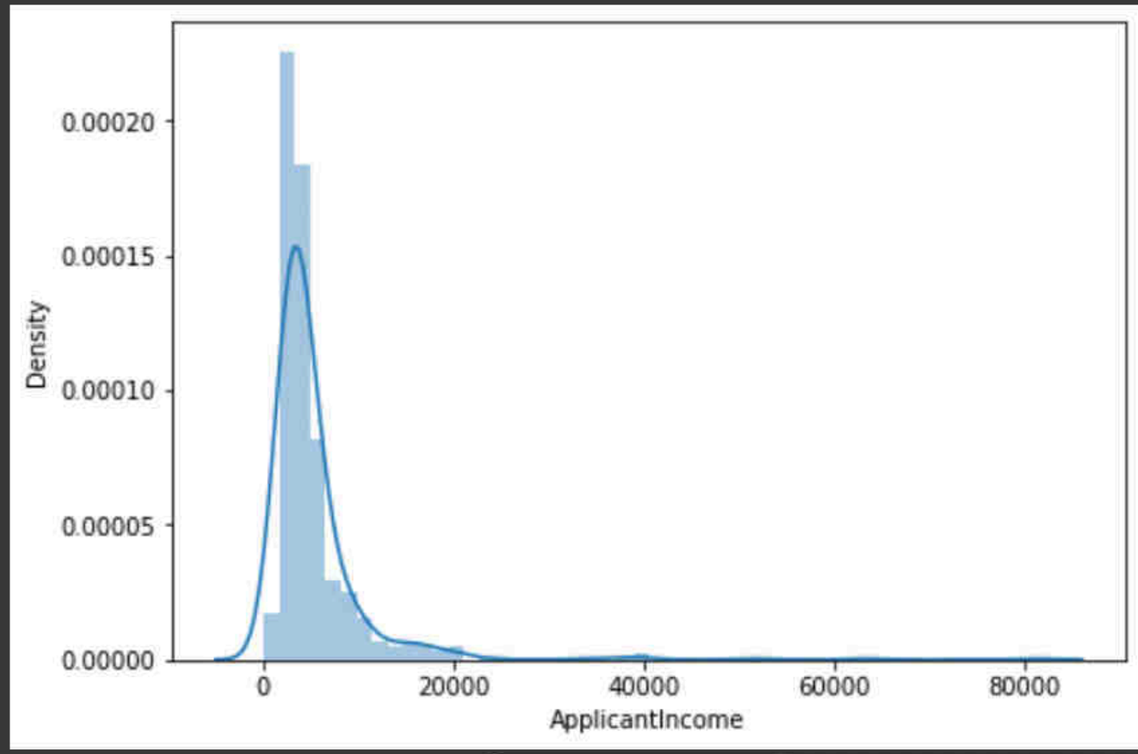


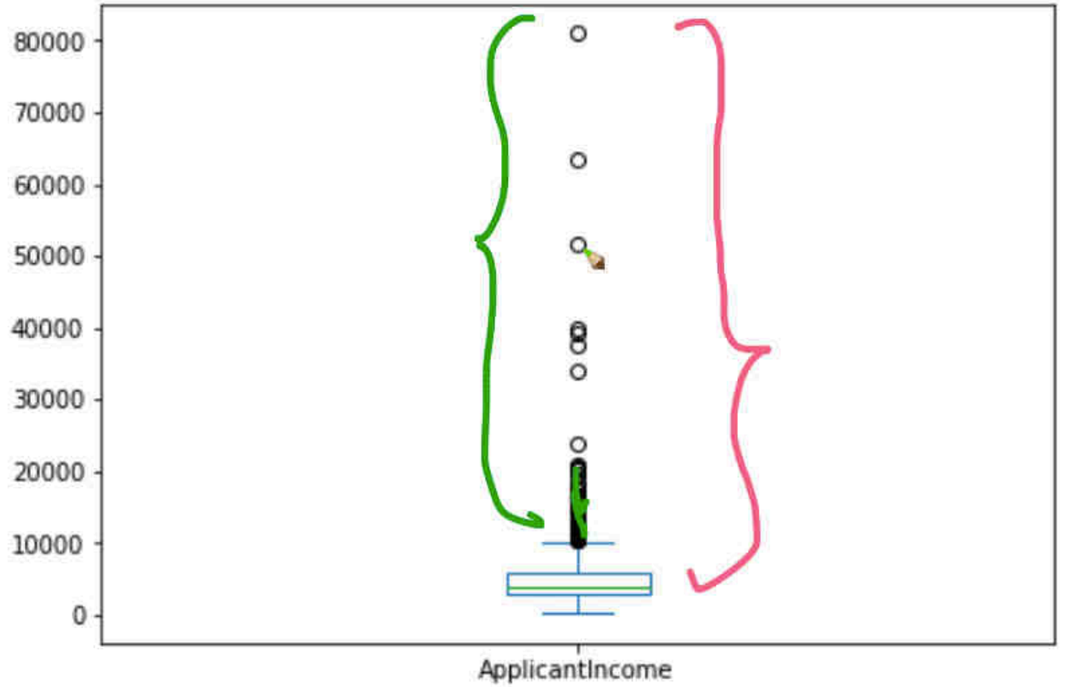

```
[ ] plt.subplot(121)
    sns.distplot(np.log(data["ApplicantIncome"]))
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help


```
[ ]  #Income of the applicant
     plt.subplot(121)
     sns.distplot(data["ApplicantIncome"])

     plt.subplot(122)
     data["ApplicantIncome"].plot.box(figsize=(16,5))
     plt.show()
```

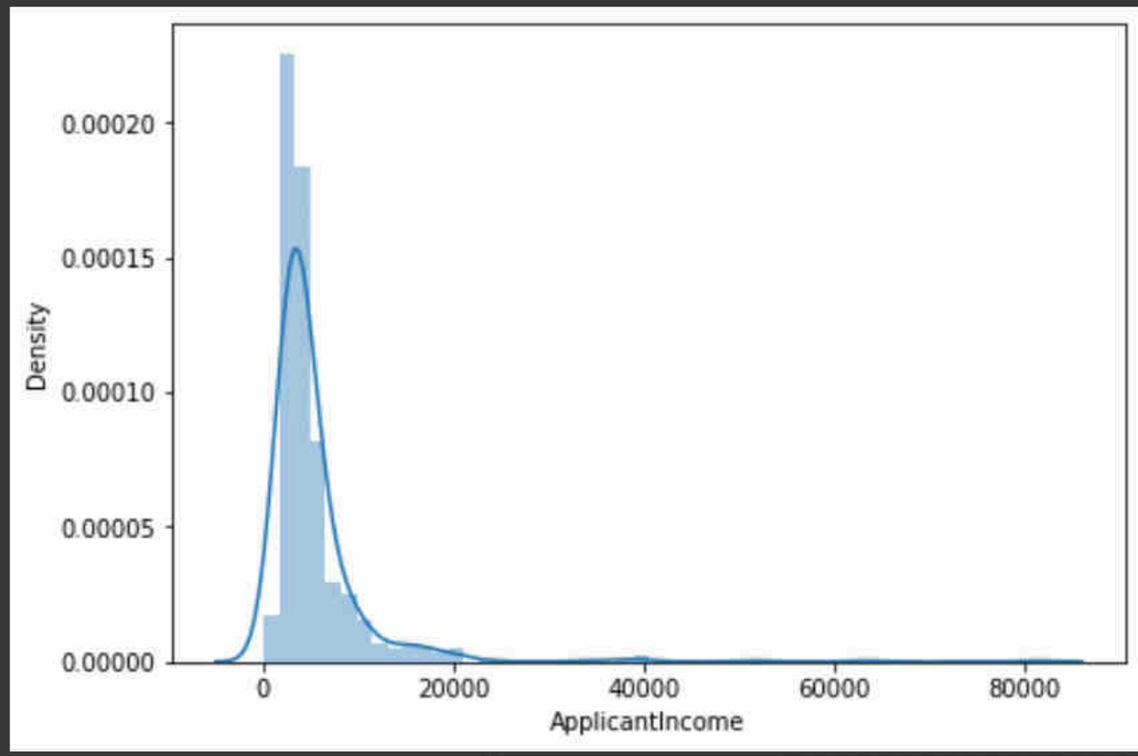
IQR ?

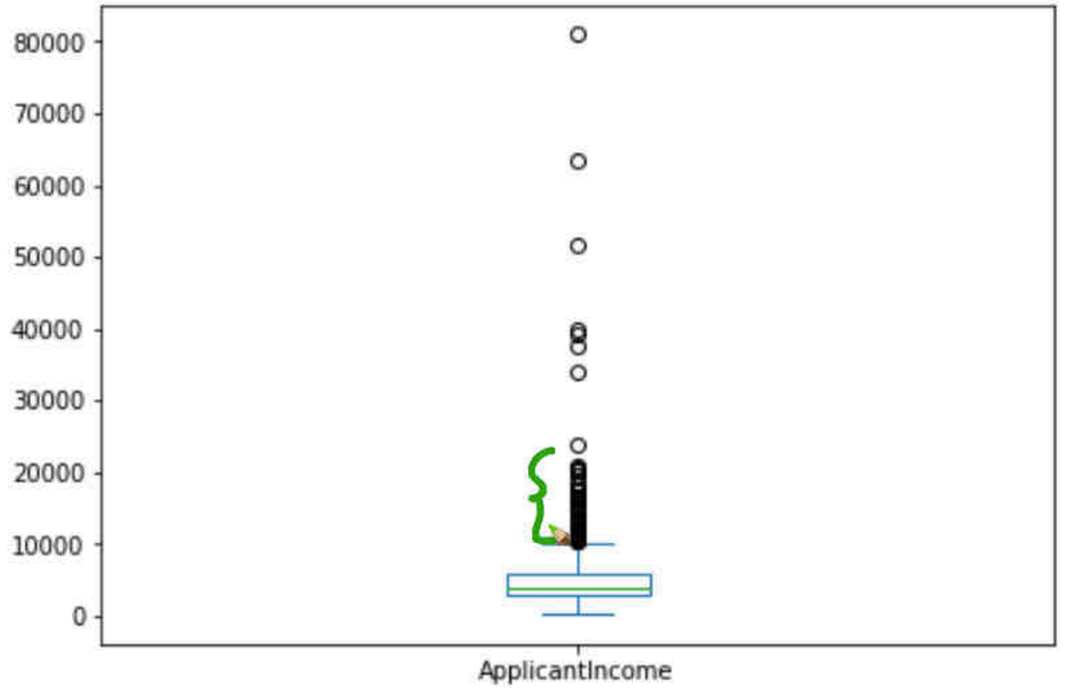




Chrome File Edit View History Bookmarks Profiles Tab Window Help

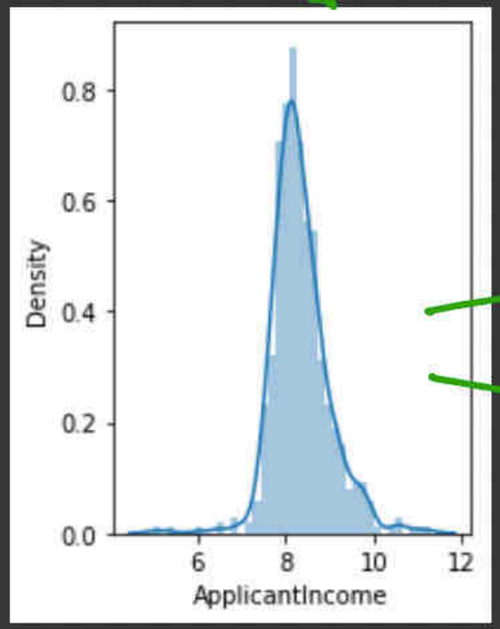

```
[ ]  #Income of the applicant
     plt.subplot(121)
     sns.distplot(data["ApplicantIncome"])

     plt.subplot(122)
     data["ApplicantIncome"].plot.box(figsize=(16,5))
     plt.show()
```





ApplicantIncome

```
plt.subplot(121)
sns.distplot(np.log(data["ApplicantIncome"]))

plt.show()
```



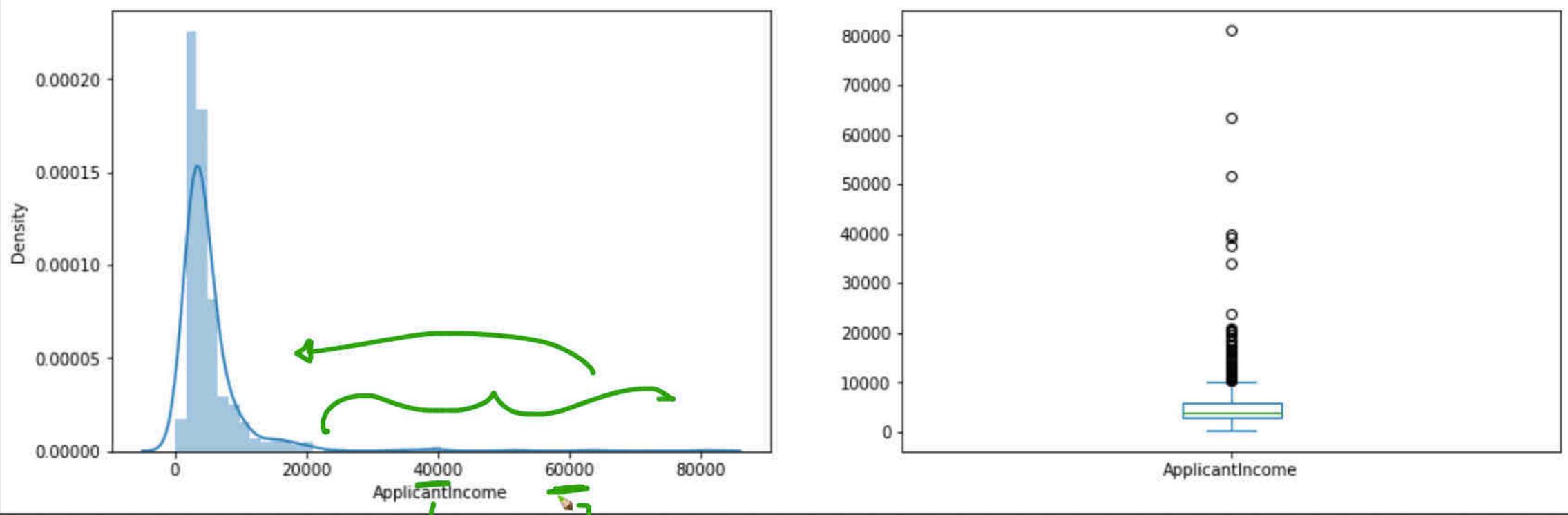
QQ plot

ks test / AD test

```
[ ] #Slice this data by Education
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:41
10.2 - Hypothesis Testing | ST... | Kruskal–Wallis one-way analy... | Anscombe's quartet 3 - Ansc... | Bias of an estimator - Wikiped... | EDA_FE.ipynb - Colaboratory | scipy.stats.kruskal — SciPy v1... | numpy.var — NumPy v1.23 Ma...
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=mg6keQsT_PL5
+ Code + Text Reconnect

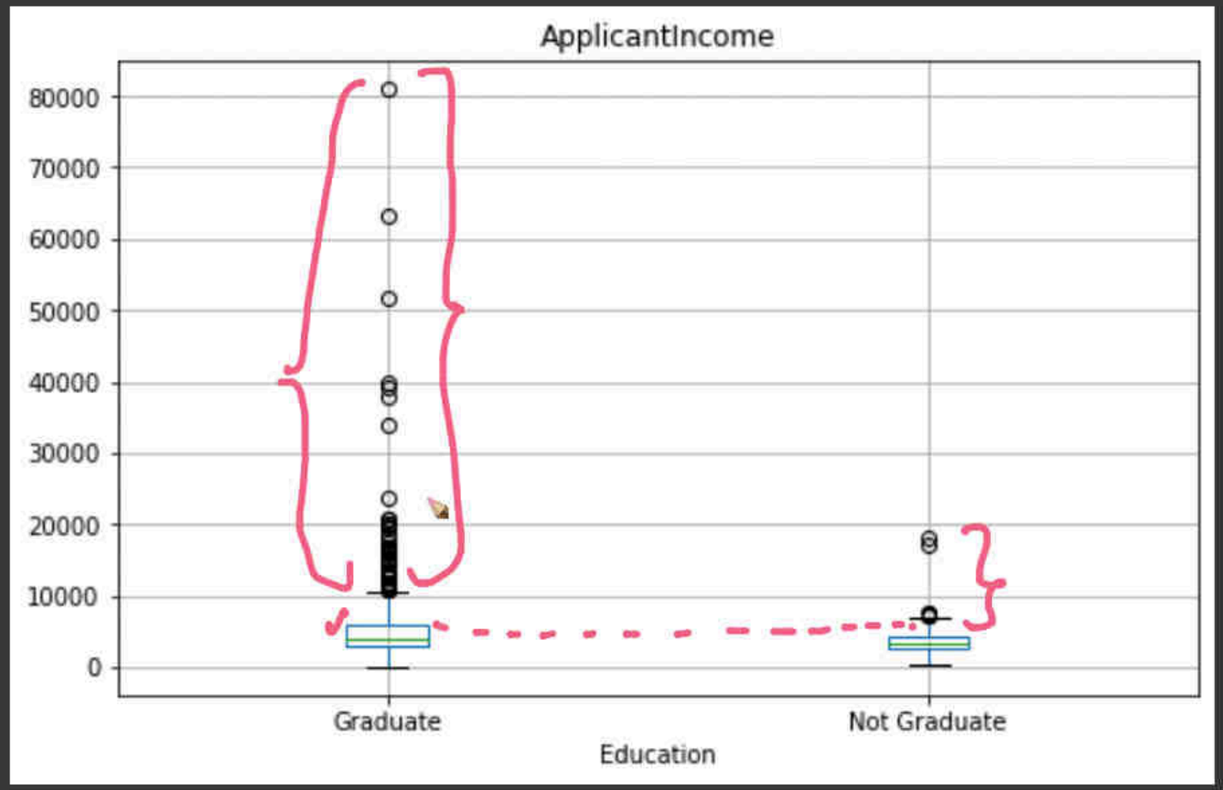
```
data["ApplicantIncome"].plot.box(figsize=(16,5))
plt.show()
```




```
plt.subplot(121)
sns.distplot(np.log(data["ApplicantIncome"]))

plt.show()
```

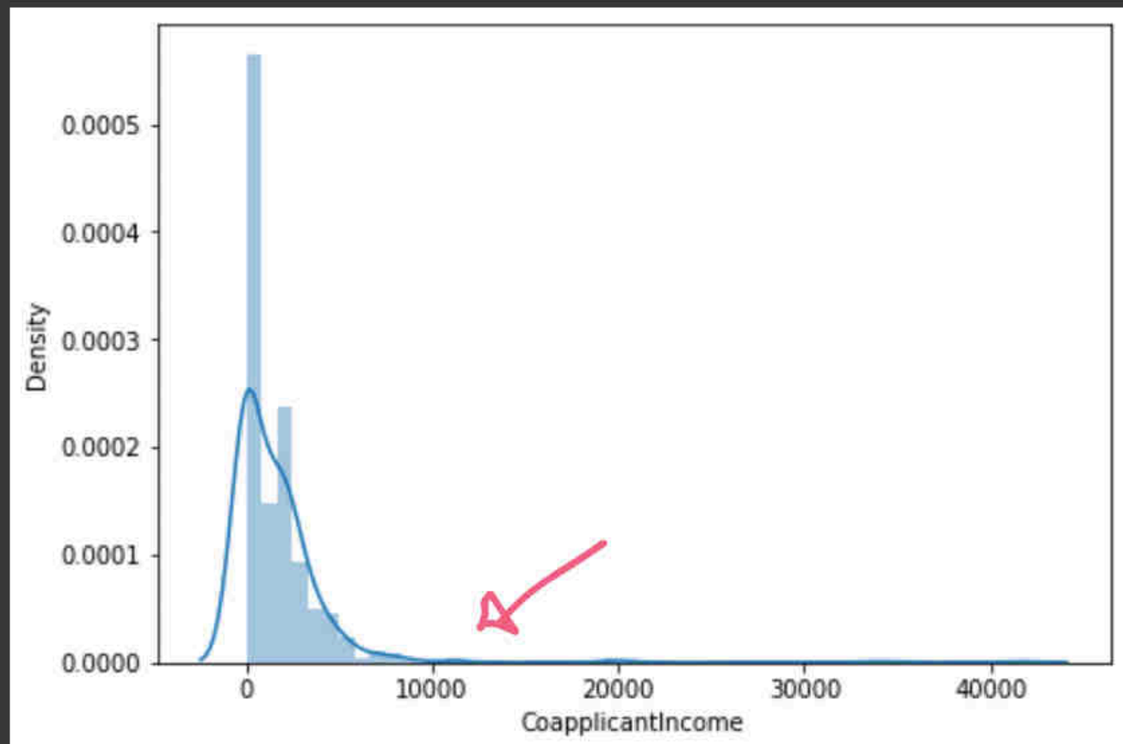
82 / 82

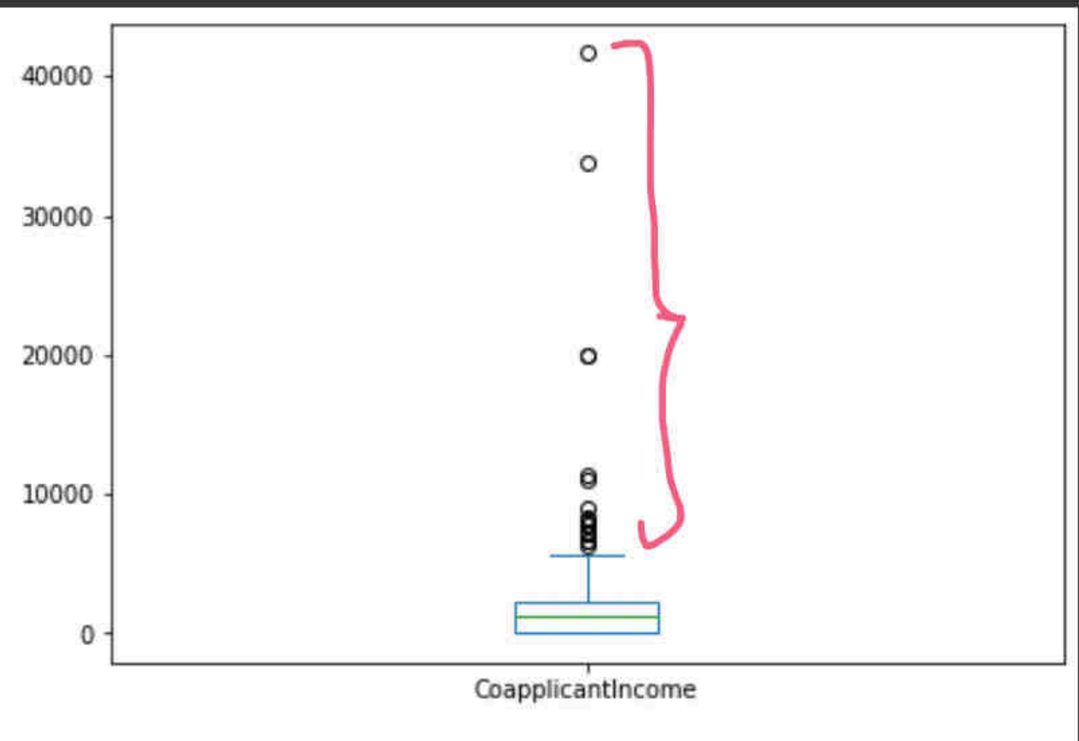
```
data.boxplot(column='ApplicantIncome', by="Education", figsize=(8,5))
plt.suptitle("")
plt.show()
```




```
[ ]  #co-applicant income
     plt.subplot(121)
```

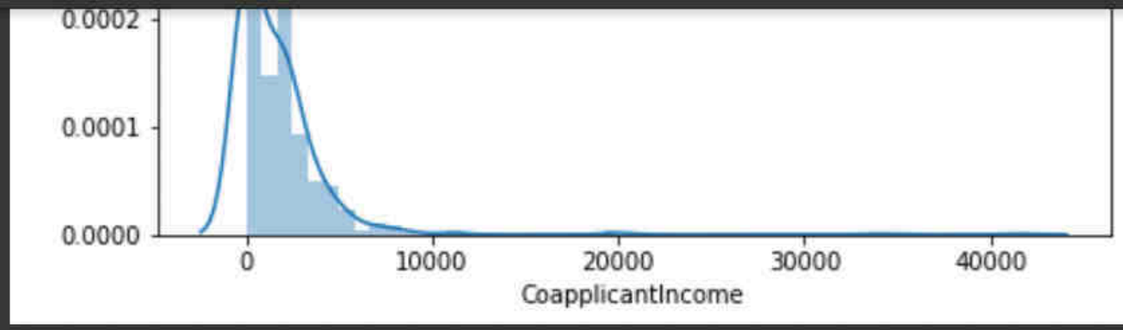
```
plt.subplot(122)
data["CoapplicantIncome"].plot.box(figsize=(16,5))
plt.show()
```

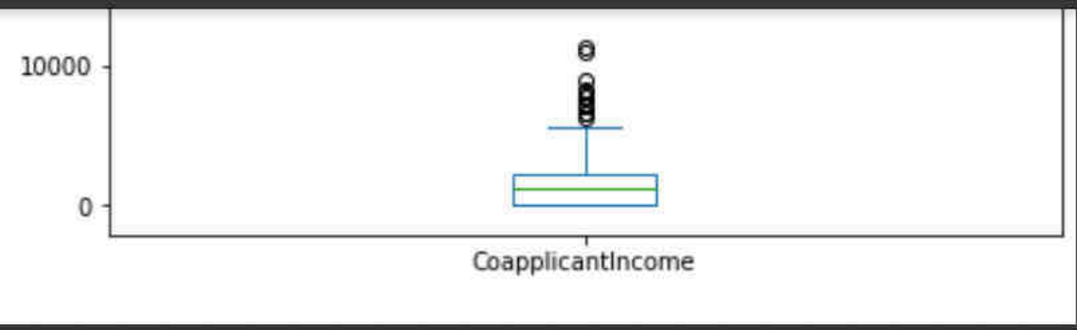





```
[ ]  #Relation between "Loan_Status" and "Income"
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:45
10.2 - Hypothesis Testing | ST... | Kruskal–Wallis one-way analy... | Anscombe's quartet 3 - Ansc... | Bias of an estimator - Wikiped... | EDA_FE.ipynb - Colaboratory | scipy.stats.kruskal — SciPy v1... | numpy.var — NumPy v1.23 Ma...
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=gwIudxInKDvf
+ Code + Text Reconnect



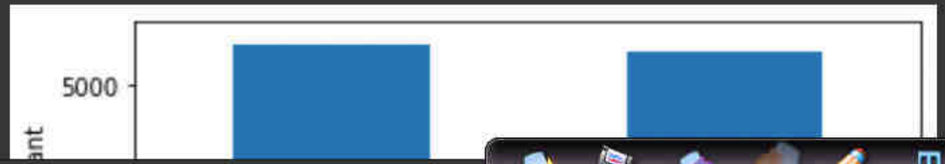



```
#Relation between "Loan_Status" and "Income"
```

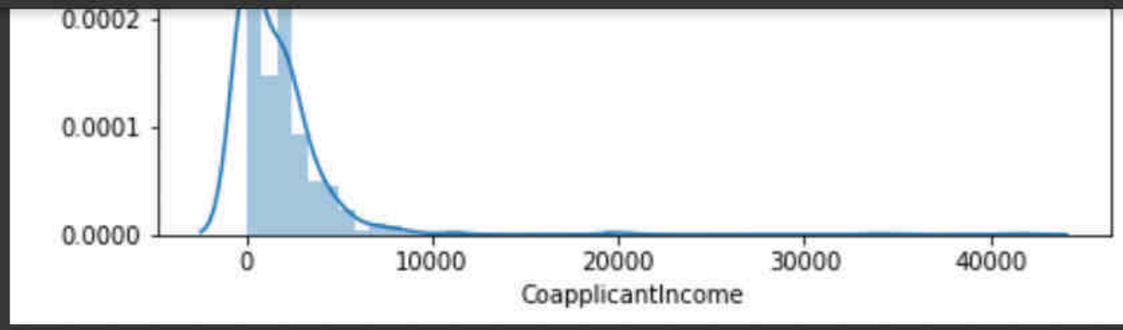
```
[ ] data.groupby("Loan_Status").mean()['ApplicantIncome']
```

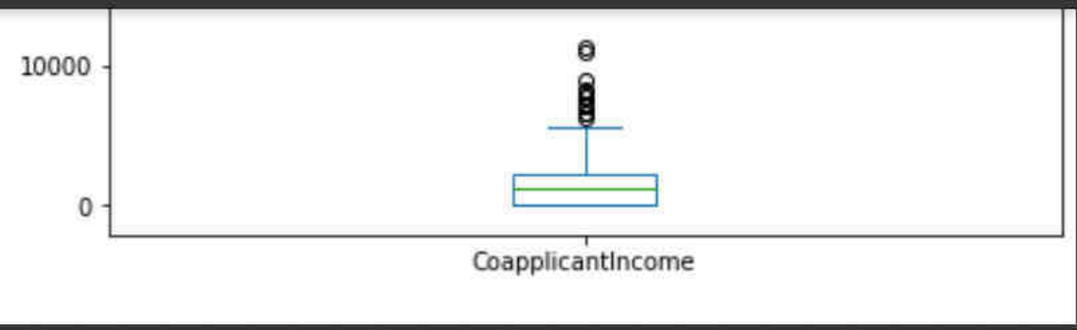
```
Loan_Status
N    5446.078125
Y    5384.068720
Name: ApplicantIncome, dtype: float64
```

```
data.groupby("Loan_Status").mean()['ApplicantIncome'].plot.bar()
plt.ylabel("Mean Income of applicant")
plt.show()
```



85 / 85



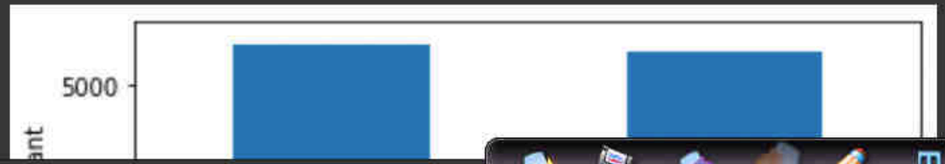



```
#Relation between "Loan_Status" and "Income"
```

```
[ ] data.groupby("Loan_Status").mean()['ApplicantIncome']
```

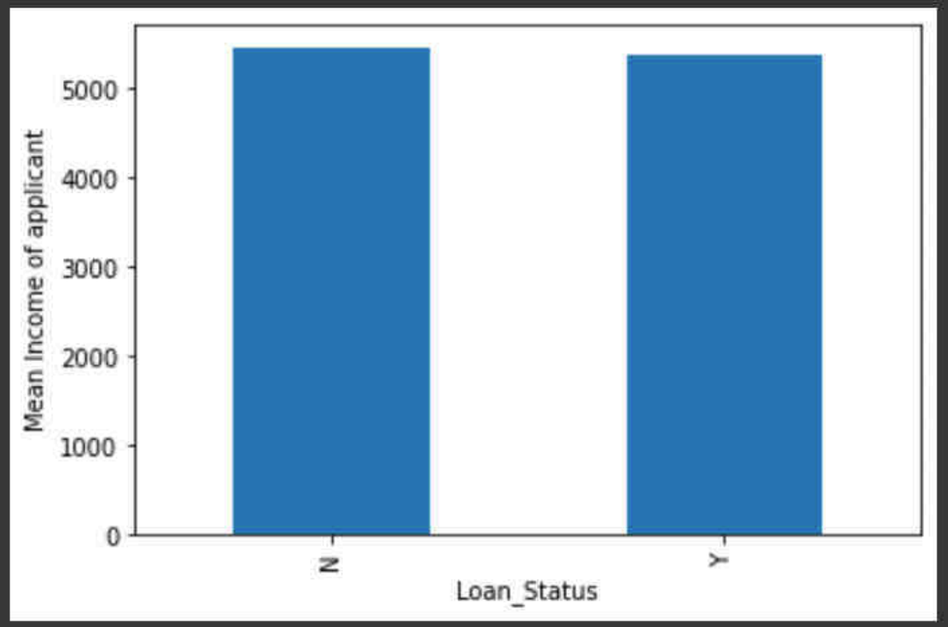
```
Loan_Status
N    5446.078125
Y    5384.068720
Name: ApplicantIncome, dtype: float64
```

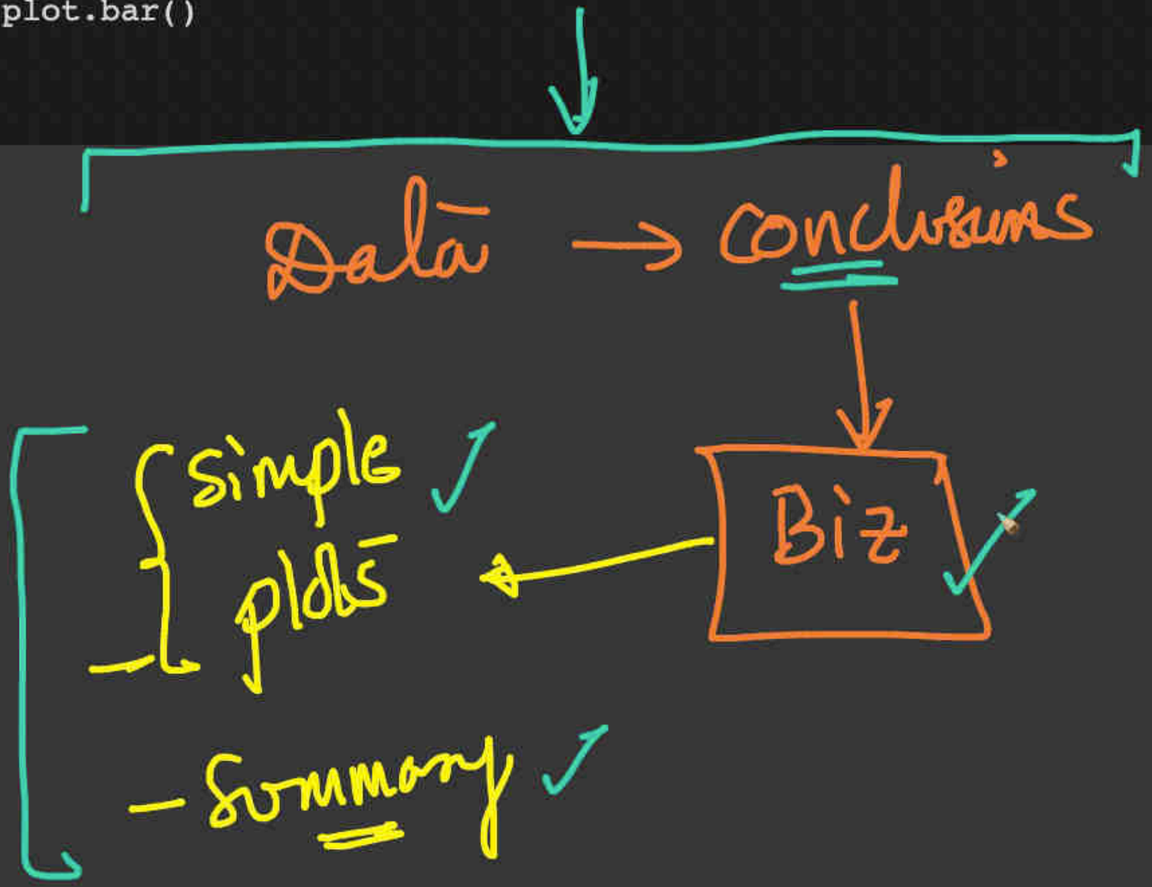
```
data.groupby("Loan_Status").mean()['ApplicantIncome'].plot.bar()
plt.ylabel("Mean Income of applicant")
plt.show()
```



Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:48
10.2 - Hypothesis Testing | ST... | Kruskal–Wallis one-way analy... | Anscombe's quartet 3 - Ansc... | Bias of an estimator - Wikiped... | EDA_FE.ipynb - Colaboratory | scipy.stats.kruskal — SciPy v1... | numpy.var — NumPy v1.23 Ma...
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=gwIudxlnKDvf
Update
+ Code + Text Reconnect

```
[ ]  data.groupby("Loan_Status").mean()['ApplicantIncome'].plot.bar()
     plt.ylabel("Mean Income of applicant")
     plt.show()
```





Simple Feature Engineering

[ ] ↳ 37 cells hidden

87 / 87

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:49

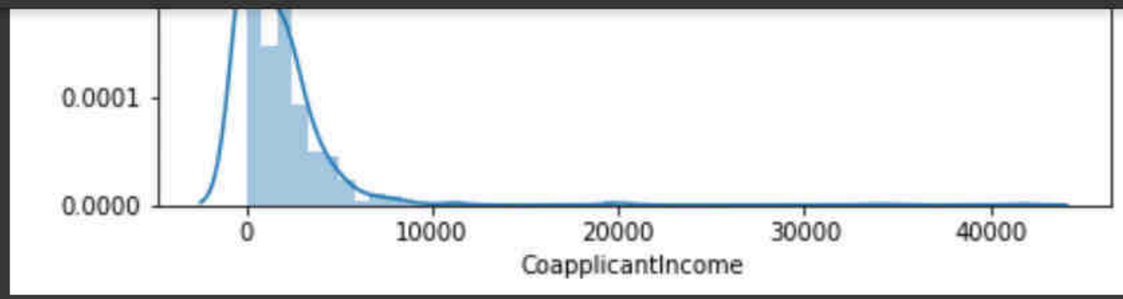
Simple Feature Engineering

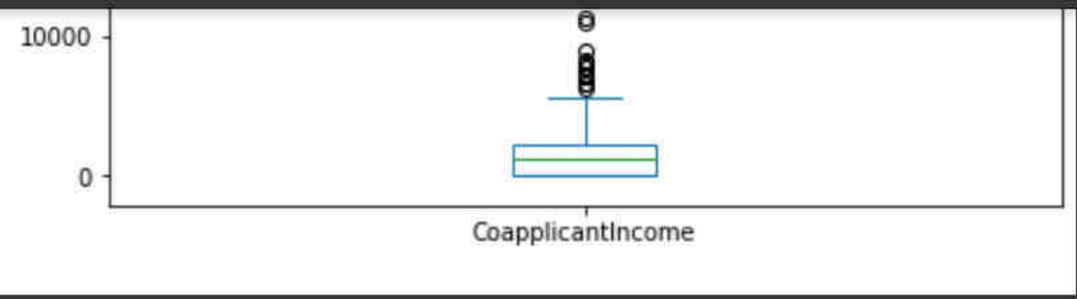
```
[ ]  # Feature binning: income
     bins=[0,2500,4000,6000, 8000, 10000, 20000, 40000, 81000]
     group=['Low','Average','medium', 'H1', 'h2', 'h3', 'h4' , 'Very high']
     data['Income_bin']= pd.cut(data['ApplicantIncome'],bins,labels=group)
```

```
[ ]  data.head()
```

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_
0	Male	No	0	Graduate	No	5849	0.0	NaN	
1	Male	Yes	1	Graduate	No	4583	1508.0	128.0	
2	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	
3	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	
4	Male	No	0	Graduate	No	6000	0.0	141.0	

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:49
10.2 - Hypothesis Testing | ST... | Kruskal–Wallis one-way analy... | Anscombe's quartet 3 - Ansc... | Bias of an estimator - Wikiped... | EDA_FE.ipynb - Colaboratory | scipy.stats.kruskal — SciPy v1... | numpy.var — NumPy v1.23 Ma...
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=sp3SJkJmKaIm
+ Code + Text Reconnect



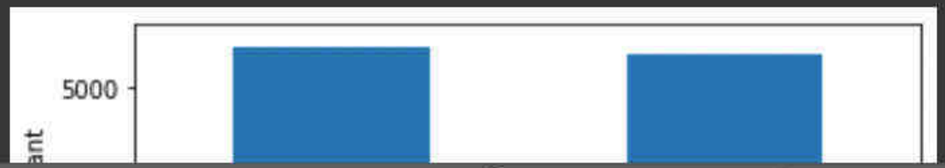



```
[ ] #Relation between "Loan_Status" and "Income"
```

```
[ ] data.groupby("Loan_Status").mean()['ApplicantIncome']
```

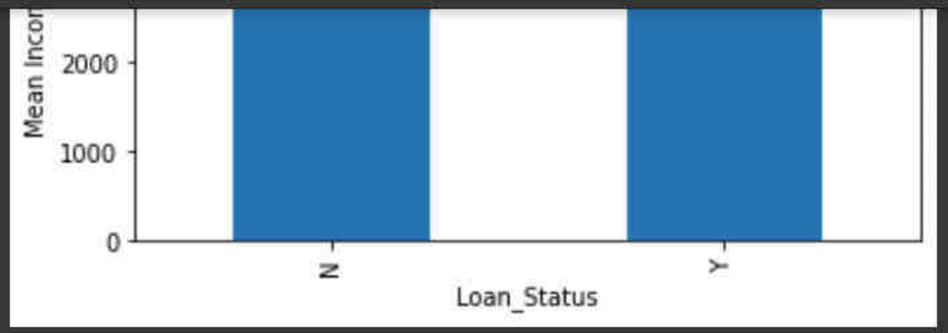
```
Loan_Status
N    5446.078125
Y    5384.068720
Name: ApplicantIncome, dtype: float64
```

```
[ ] data.groupby("Loan_Status").mean()['ApplicantIncome'].plot.bar()
    plt.ylabel("Mean Income of applicant")
    plt.show()
```



89 / 89

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:50



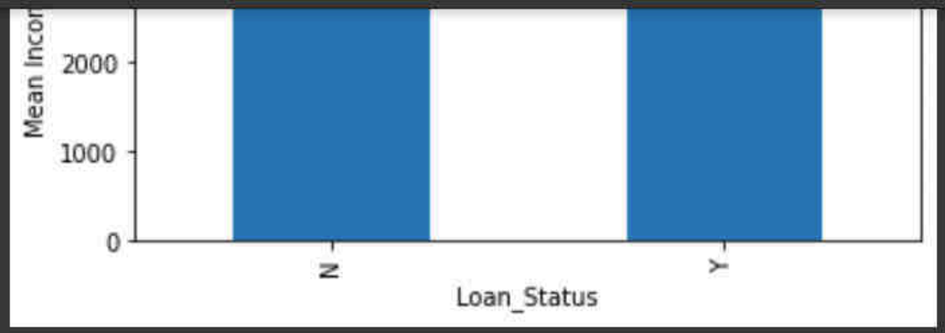
Transformation

Simple Feature Engineering

```
# Feature binning: income
bins=[0,2500,4000,6000, 8000, 10000, 20000, 40000, 81000]
group=['Low','Average','medium', 'H1', 'h2', 'h3', 'h4' , 'Very high']
data['Income_bin']= pd.cut(data['ApplicantIncome'],bins,labels=group)
```

```
[ ]  data.head()
```

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_
0	Male	No	0	Graduate	No	5849	0.0	NaN	
1	Male	Yes	1	Graduate	No	4583	1508.0	128.0	



0-2500 →

Simple Feature Engineering

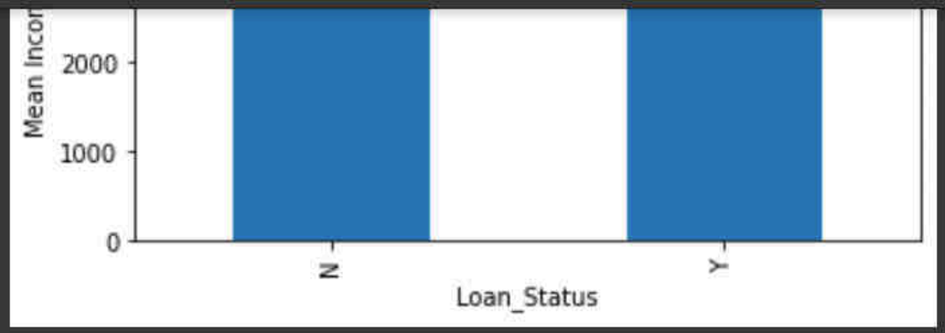
```
# Feature binning: income
bins=[0,2500,4000,6000, 8000, 10000, 20000, 40000, 81000]
group=['Low','Average','medium', 'H1', 'h2', 'h3', 'h4' , 'Very high']
data['Income_bin']= pd.cut(data['ApplicantIncome'],bins,labels=group)
```

```
[ ]  data.head()
```

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_
0	Male	No	0	Graduate	No	5849	0.0	NaN	
1	Male	Yes	1	Graduate	No	4583	1508.0	128.0	

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:51

10.2 - Hypothesis Testing | ST... | Kruskal–Wallis one-way analy... | Anscombe's quartet 3 - Ansc... | Bias of an estimator - Wikiped... | EDA_FE.ipynb - Colaboratory | scipy.stats.kruskal — SciPy v1... | numpy.var — NumPy v1.23 Ma...
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=19FtHN8LKZS_



Simple Feature Engineering

Domain Knowledge

```
# Feature binning: income
bins=[0,2500,4000,6000, 8000, 10000, 20000, 40000, 81000]
group=['Low','Average','medium', 'H1', 'h2', 'h3', 'h4' , 'Very high']
data['Income_bin']= pd.cut(data['ApplicantIncome'],bins,labels=group)
```

```
[ ]  data.head()
```

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_
0	Male	No	0	Graduate	No	5849	0.0	NaN	
1	Male	Yes	1	Graduate	No	4583	1508.0	128.0	

income - bins

0 — 81,000

→ domain-specific

→ 5 bins: 0-20 21-40 41-60 61-80 81-100

percentiles

→ ML: decision trees (automatically bin)

(later)

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:54
10.2 - Hypothesis Testing | ST... | Kruskal–Wallis one-way analy... | Anscombe's quartet 3 - Ansc... | Bias of an estimator - Wikiped... | EDA_FE.ipynb - Colaboratory | scipy.stats.kruskal — SciPy v1... | numpy.var — NumPy v1.23 Ma...
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=19FtHN8LKZS_
+ Code + Text Reconnect

Simple Feature Engineering

```
# Feature binning: income
bins=[0,2500,4000,6000, 8000, 10000, 20000, 40000, 81000]
group=['Low','Average','medium', 'H1', 'h2', 'h3', 'h4' , 'Very high']
data['Income_bin']= pd.cut(data['ApplicantIncome'],bins,labels=group)
```

```
[ ]  data.head()
```

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_
0	Male	No	0	Graduate	No	5849	0.0	NaN	
1	Male	Yes	1	Graduate	No	4583	1508.0	128.0	
2	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	
3	Male	Yes	0	Not	No	2583	2358.0	120.0	

94 / 94

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Incomes

```
[ ] #observed
    pd.crosstab(data["Income_bin"],data["Loan_Status"])
```

Loan_Status	N	Y
Income_bin		
Low	34	74
Average	67	159
medium	45	98
H1	20	34
h2	9	22
h3	13	27
h4	3	6
Very high	1	2

Chrome File Edit View History Bookmarks Profiles Tab Window Help
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | scipy.stats.kruskal | numpy.var — NumPy | pandas.cut — panda | pandas.qcut — panda
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=19FtHN8LKZS_
Update

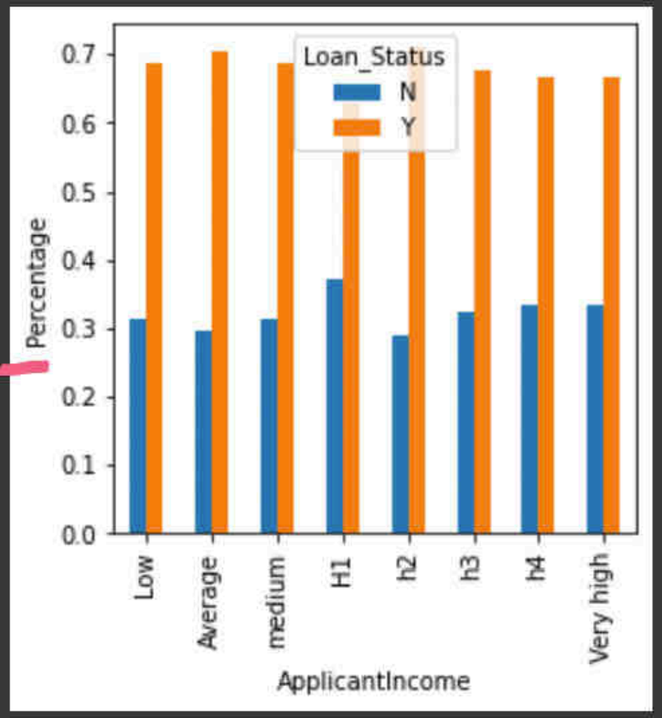
Incomes

```
[ ] #observed
    pd.crosstab(data["Income_bin"],data["Loan_Status"])
```

Loan_Status	N	Y
Income_bin		
Low	34	74
Average	67	159
medium	45	98
H1	20	34
h2	9	22
h3	13	27
h4	3	6
Very high	1	2

96 / 96

```
    plt.xlabel("ApplicantIncome")
    plt.ylabel("Percentage")
    plt.show()

    #It can be inferred that Applicant income does not affect the chances of loan approval. Which seems wrong as per ou
```



$P(L=Y \mid Low)$

$P(L=N \mid Low)$

```
[ ]  #co-appplicant income
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help
10:59
10.2 - Hypothesis Te | Kruskal–Wallis one-w | Anscombe's quartet | Bias of an estimator | EDA_FE.ipynb - Colab | scipy.stats.kruskal | numpy.var — NumPy | pandas.cut — panda | pandas.qcut — panda
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=19FtHN8LKZS_
Update
+ Code + Text
Reconnect

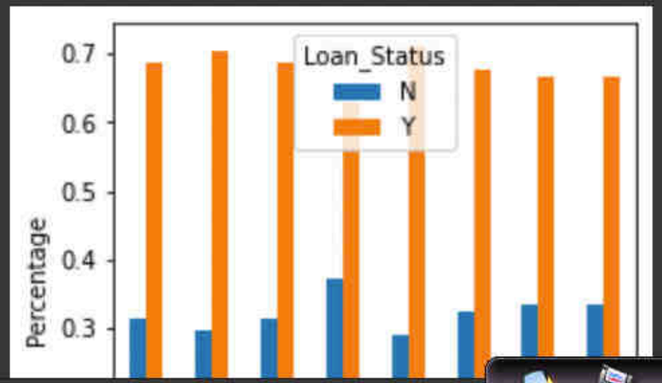
```
Income_bin.div(Income_bin.sum(axis=1),axis=0).plot(kind="bar",figsize=(4,4))
plt.xlabel("ApplicantIncome")
plt.ylabel("Percentage")
plt.show()

#It can be inferred that Applicant income does not affect the chances of loan approval. Which seems wrong as per ou
```


```
[ ]  #co-appplicant income
```

98 / 98

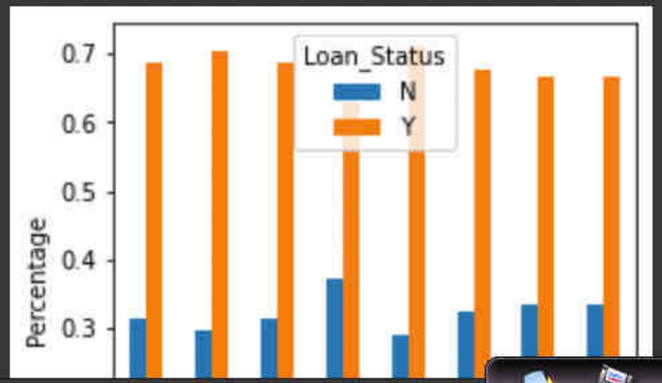
h2	9	22
h3	13	27
h4	3	6
Very high	1	2

```
[ ]  Income_bin = pd.crosstab(data["Income_bin"],data["Loan_Status"])

     Income_bin.div(Income_bin.sum(axis=1),axis=0).plot(kind="bar",figsize=(4,4))
     plt.xlabel("ApplicantIncome")
     plt.ylabel("Percentage")
     plt.show()

     #It can be inferred that Applicant income does not affect the chances of loan approval. Which seems wrong as per ou
```

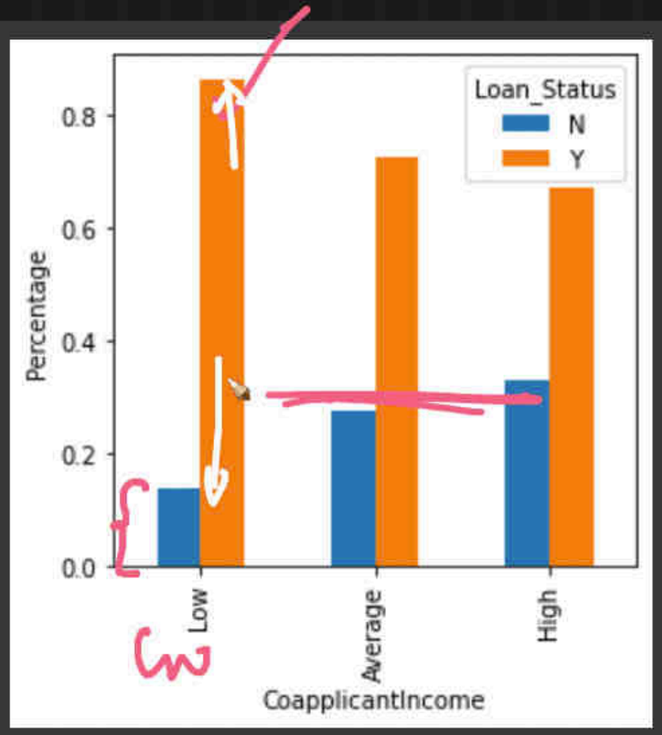


h2	9	22
h3	13	27
h4	3	6
Very high	1	2

```
[ ]  Income_bin = pd.crosstab(data["Income_bin"],data["Loan_Status"])

     Income_bin.div(Income_bin.sum(axis=1),axis=0).plot(kind="bar",figsize=(4,4))
     plt.xlabel("ApplicantIncome")
     plt.ylabel("Percentage")
     plt.show()

     #It can be inferred that Applicant income does not affect the chances of loan approval. Which seems wrong as per ou
```



```
CoapplicantIncome_Bin = pd.crosstab(data["CoapplicantIncome_bin"],data["Loan_Status"])
CoapplicantIncome_Bin.div(CoapplicantIncome_Bin.sum(axis = 1),axis=0).plot(kind='bar',figsize=(4,4))
plt.xlabel("CoapplicantIncome")
plt.ylabel("Percentage")
plt.show()

## What's the problem here? Why co-applicant having low income is getting maximum loan approved?
```



Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:06
10.2 - Hypothesis | Kruskal–Wallis one | Anscombe's quart | Bias of an estimat | EDA_FE.ipynb - C | scipy.stats.kruska | numpy.var — Num | pandas.cut — pan | pandas.qcut — pa | pandas.DataFrame
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=TarGSsXDG8Tp
Update
+ Code + Text
Reconnect

```
# only numeric features
```

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

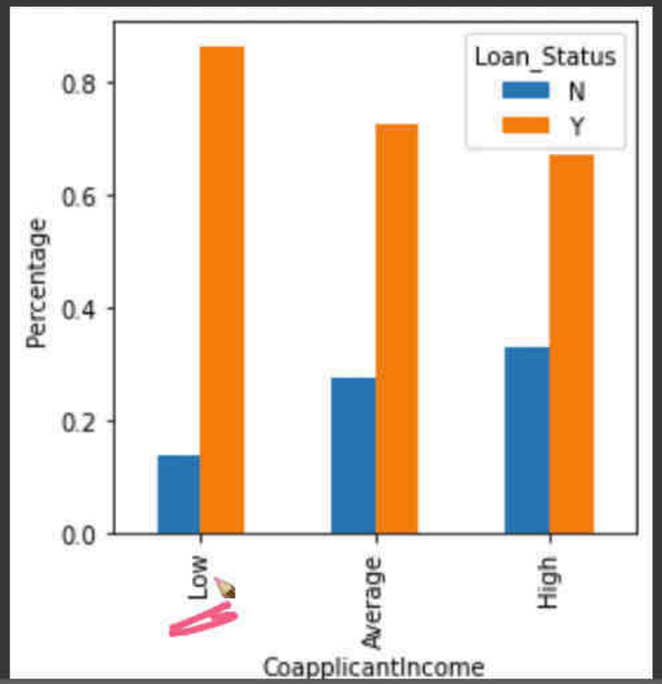
low
0 – 1000

```
[ ] # catgeorical features
    data.describe(include = ['object'])
```

	Gender	Married	Dependents	Education	Self_Employed	Property_Area	Loan_Status
count	601	611	599	614	582	614	614
unique	2	2	4	2	2	3	2

102 / 102

```
CoapplicantIncome_Bin = pd.crosstab(data["CoapplicantIncome_bin"],data["Loan_Status"])
CoapplicantIncome_Bin.div(CoapplicantIncome_Bin.sum(axis = 1),axis=0).plot(kind='bar',figsize=(4,4))
plt.xlabel("CoapplicantIncome")
plt.ylabel("Percentage")
plt.show()

## What's the problem here? Why co-applicant having low income is getting maximum loan approved?
```



Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:08
10.2 - Hypothesis | Kruskal–Wallis one | Anscombe's quart | Bias of an estimat | EDA_FE.ipynb - C | scipy.stats.kruske | numpy.var — Num | pandas.cut — pan | pandas.qcut — pa | pandas.DataFrame
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=TarGSsXDG8Tp
Update
+ Code + Text
Reconnect

613

```
[ ] data['CoapplicantIncome'].value_counts().head()
```

```
0.0       273
2500.0      5
2083.0      5
1666.0      5
2250.0      3
Name: CoapplicantIncome, dtype: int64
```

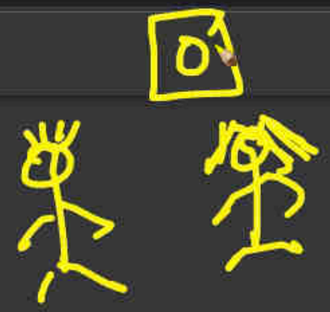
```
[ ] # New feature:  total household income
    data["TotalIncome"] = data["ApplicantIncome"] + data["CoapplicantIncome"]
```

```
[ ] bins = [0,3000,5000,8000,81000]
    group = ['Low','Average','High','Very High']
    data["TotalIncome_bin"] = pd.cut(data["TotalIncome"],bins,labels=group)
```

```
[ ] pd.crosstab(data["TotalIncome_bin"], data["Loan_Status"])
```

104 / 104

```
2500.0       5
2083.0       5
1666.0       5
2250.0       3
Name: CoapplicantIncome, dtype: int64
```



```
# New feature: total household income
data["TotalIncome"] = data["ApplicantIncome"] + data["CoapplicantIncome"]
```

```
bins = [0,3000,5000,8000,81000]
group = ['Low','Average','High','Very High']
data["TotalIncome_bin"] = pd.cut(data["TotalIncome"],bins,labels=group)
```

```
pd.crosstab(data["TotalIncome_bin"], data["Loan_Status"])
```

Loan_Status	N	Y
TotalIncome_bin		
Low	20	27
Average	69	154
High	61	151

```
2500.0      5
2083.0      5
1666.0      5
2250.0      3
Name: CoapplicantIncome, dtype: int64
```

```
# New feature: total household income
data["TotalIncome"] = data["ApplicantIncome"] + data["CoapplicantIncome"]
```

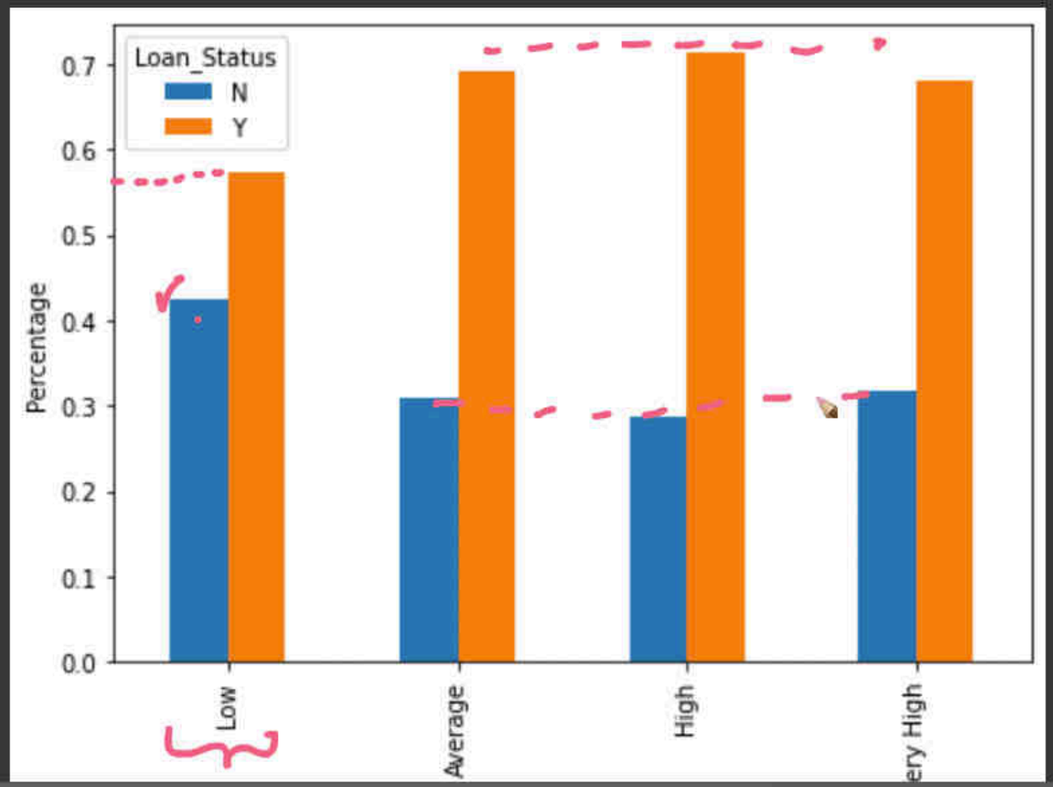
```
bins = [0,3000,5000,8000,81000]
group = ['Low','Average','High','Very High']
data["TotalIncome_bin"] = pd.cut(data["TotalIncome"],bins,labels=group)
```

```
pd.crosstab(data["TotalIncome_bin"], data["Loan_Status"])
```

Loan_Status	N	Y
TotalIncome_bin		
Low	20	27
Average	69	154
High	61	151

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:11
10.2 - Hypothesis | Kruskal–Wallis one | Anscombe's quart | Bias of an estimat | EDA_FE.ipynb - C | scipy.stats.kruska | numpy.var — Num | pandas.cut — pan | pandas.qcut — pa | pandas.DataFrame
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=oDO5pryhMK21
+ Code + Text Reconnect

```
plt.xlabel("TotalIncome")
plt.ylabel("Percentage")
plt.show()

# Observation: We can see that Proportion of loans getting approved for
# applicants having low Total_Income is very less as compared to that of applicants
# with Average, High and Very High Income.
```



107 / 107

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:14

10.2 - Hypothesis | Kruskal–Wallis one | Anscombe's quart | Bias of an estimat | EDA_FE.ipynb - C | scipy.stats.krusk | numpy.var — Num | pandas.cut — pan | pandas.qcut — pa | pandas.DataFrame

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=hIyzygASMSB_ Update

+ Code + Text Reconnect


```
[ ]  data = data.drop(["Income_bin","CoapplicantIncome_bin","TotalIncome_bin"],axis=1)
```

Loan Amount and Loan Term

① binning

② combine features

```
[ ]  data['Loan_Amount_Term'].value_counts()
```

```
360.0    512
180.0     44
480.0     15
300.0     13
240.0      4
84.0       4
120.0      3
60.0       2
36.0       2
12.0       1
Name: Loan_Amount_Term, dtype: int64
```

```
[ ]  data['Loan_Amount_Term'] = (data['Loan_Amount_Term']/12).astype('float')
```

108 / 108

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:15

10.2 - Hypothesis | Kruskal–Wallis one | Anscombe's quart | Bias of an estimat | EDA_FE.ipynb - C | scipy.stats.kruska | numpy.var — Num | pandas.cut — pan | pandas.qcut — pa | pandas.DataFrame

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=hIyzygASMSB_ Update

+ Code + Text Reconnect

Loan Amount and Loan Term

```
[ ] data['Loan_Amount_Term'].value_counts()
```

```
360.0    512
180.0     44
480.0     15
300.0     13
240.0      4
84.0       4
120.0      3
60.0       2
36.0       2
12.0       1
Name: Loan_Amount_Term, dtype: int64
```

```
[ ] data['Loan_Amount_Term'] = (data['Loan_Amount_Term']/12).astype('float')
```

```
[ ] sns.countplot(x='Loan_Amount_Term', data=data)
    plt.xlabel("Term in years")
    plt.show()
```

109 / 109

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:16

10.2 - Hypothesis | Kruskal–Wallis one | Anscombe's quart | Bias of an estimat | EDA_FE.ipynb - C | scipy.stats.kruska | numpy.var — Num | pandas.cut — pan | pandas.qcut — pa | pandas.DataFrame

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=hIyzygASMSB_ Update

+ Code + Text Reconnect

Loan Amount and Loan Term

```
[ ] data['Loan_Amount_Term'].value_counts()
```

```
360.0    512
180.0     44
480.0     15
300.0     13
240.0      4
84.0       4
120.0      3
60.0       2
36.0       2
12.0       1
Name: Loan_Amount_Term, dtype: int64
```

```
[ ] data['Loan_Amount_Term'] = (data['Loan_Amount_Term']/12).astype('float')
```

```
[ ] sns.countplot(x='Loan_Amount_Term', data=data)
    plt.xlabel("Term in years")
    plt.show()
    # Observation: We can clearly see that more than 90% of the loans were applied for 30 years
```

110 / 110

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:16

10.2 - Hypothesis | Kruskal–Wallis one | Anscombe's quart | Bias of an estimat | EDA_FE.ipynb - C | scipy.stats.kruska | numpy.var — Num | pandas.cut — pan | pandas.qcut — pa | pandas.DataFrame

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=hIyzygASMSB_ Update

+ Code + Text Reconnect

Loan Amount and Loan Term

```
[ ] data['Loan_Amount_Term'].value_counts()
```

```
360.0    512
180.0     44
480.0     15
300.0     13
240.0      4
84.0       4
120.0      3
60.0       2
36.0       2
12.0       1
Name: Loan_Amount_Term, dtype: int64
```

```
[ ] data['Loan_Amount_Term'] = (data['Loan_Amount_Term']/12).astype('float')
```

```
[ ] sns.countplot(x='Loan_Amount_Term', data=data)
    plt.xlabel("Term in years")
    plt.show()
```

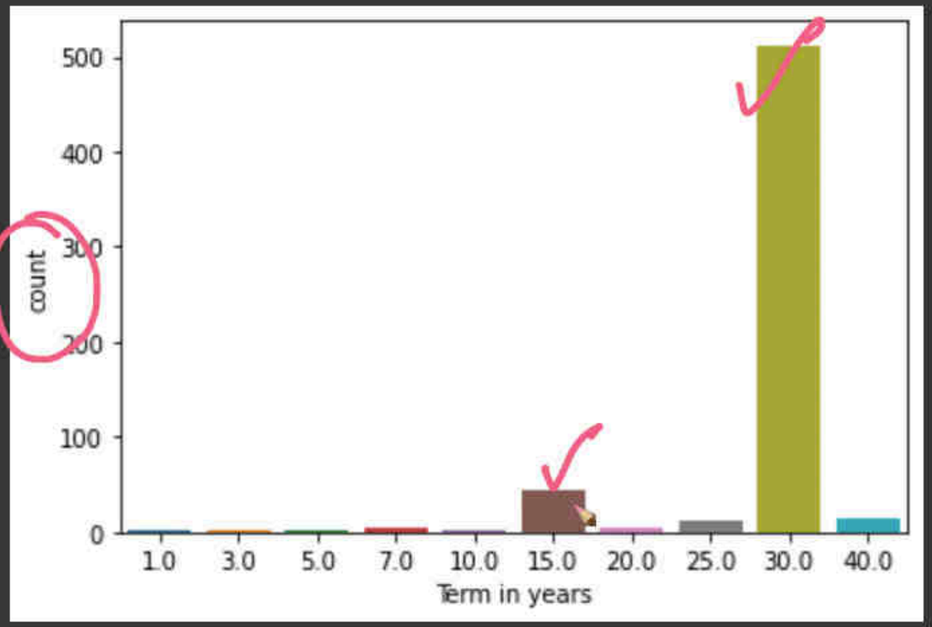
111 / 111

```
480.0     15
300.0     13
240.0      4
84.0       4
120.0      3
60.0       2
36.0       2
12.0       1
Name: Loan_Amount_Term, dtype: int64
```

```
[ ] data['Loan_Amount_Term'] = (data['Loan_Amount_Term']/12).astype('float')
```

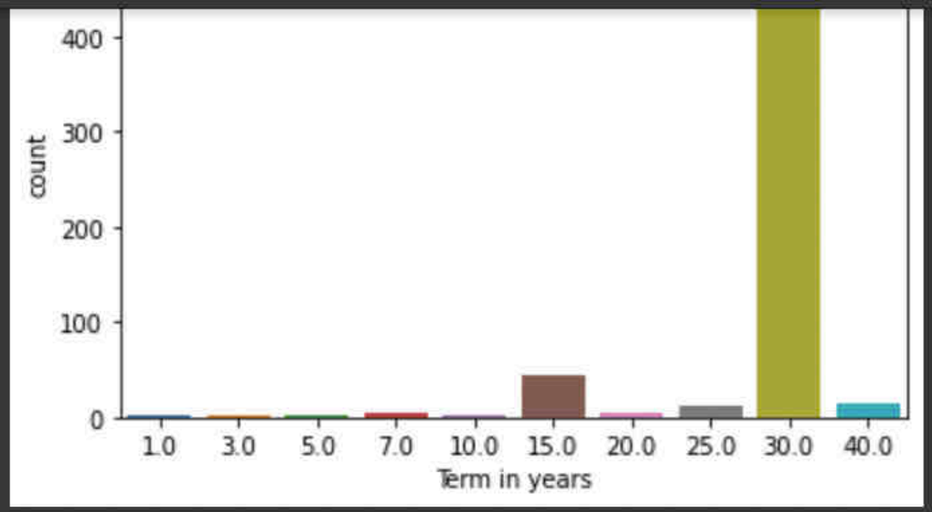
```
[ ] sns.countplot(x='Loan_Amount_Term', data=data)
    plt.xlabel("Term in years")
    plt.show()
    # Observation: We can clearly see that more than 90% of the loans were applied for 30 years.
```

```
[ ]  sns.countplot(x='Loan_Amount_Term', data=data)
     plt.xlabel("Term in years")
     plt.show()
     # Observation: We can clearly see that more than 90% of the loans were applied for 30 years.
```




```
[ ]  ### Distribution of "LoanAmount" variable :
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:18
10.2 - Hypothesis | Kruskal–Wallis one | Anscombe's quart | Bias of an estimat | EDA_FE.ipynb - C | scipy.stats.krusk | numpy.var — Num | pandas.cut — pan | pandas.qcut — pa | pandas.DataFrame
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=hIyzygASMSB_
Update
+ Code + Text Reconnect

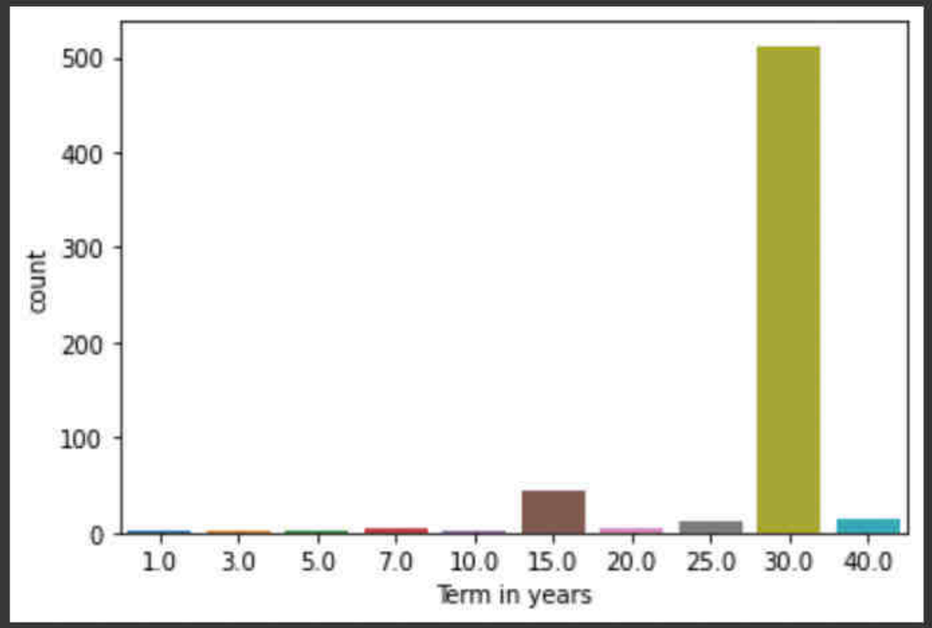


emi = f(amount, dur, IR)

```
[ ]  ### Distribution of "LoanAmount" variable :

     plt.figure(figsize=(16,5))
     plt.subplot(121)
     sns.distplot(data['LoanAmount']);

     plt.subplot(122)
     sns.boxplot(data=data, x='Loan_Status', y = 'LoanAmount')

     plt.show()
```

114 / 114

```
[ ]  sns.countplot(x='Loan_Amount_Term', data=data)
     plt.xlabel("Term in years")
     plt.show()
     # Observation: We can clearly see that more than 90% of the loans were applied for 30 years.
```



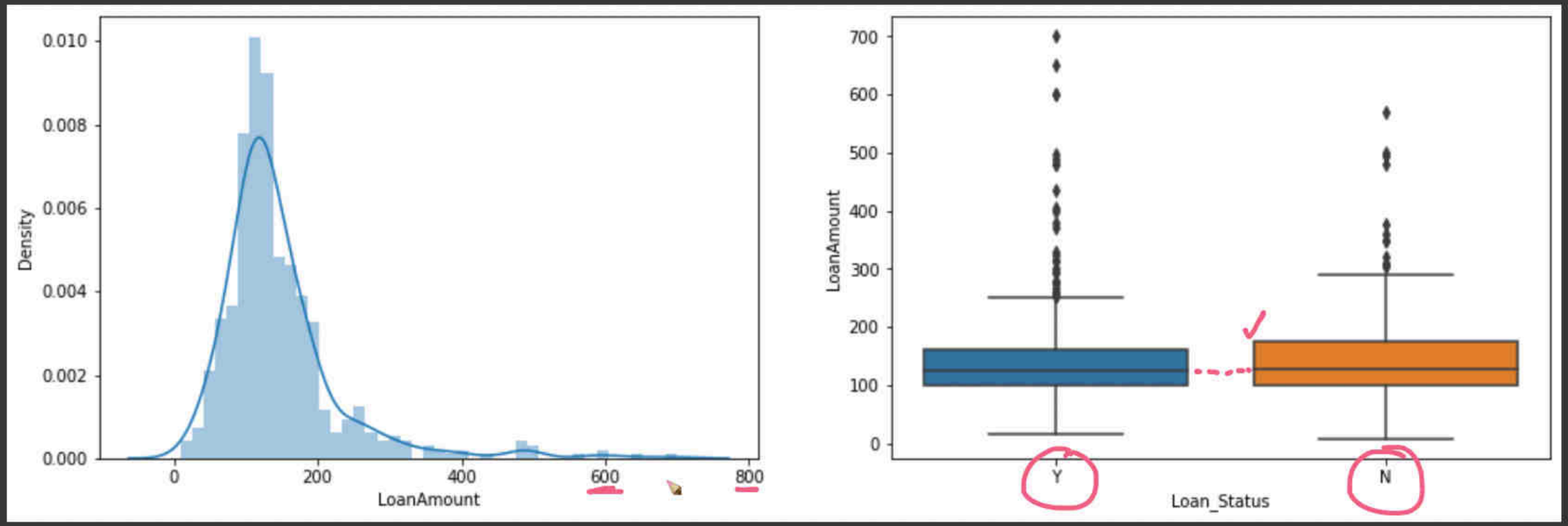
emi ≈ amount / #months ignore IR

emi → VS MI

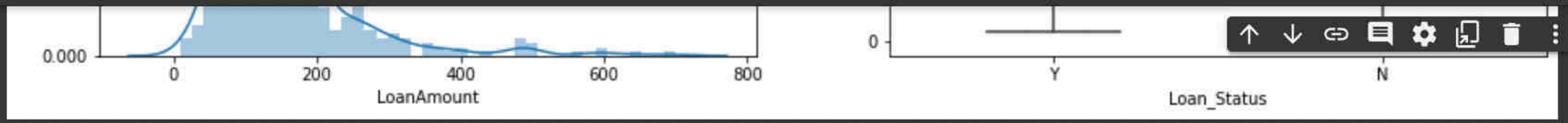
emi / MI → Y/N

```
[ ]  ### Distribution of "LoanAmount" variable :

     plt.figure(figsize=(16,5))
```

```
plt.subplot(121)
sns.distplot(data['LoanAmount']);

plt.subplot(122)
sns.boxplot(data=data, x='Loan_Status', y = 'LoanAmount')

plt.show()
```

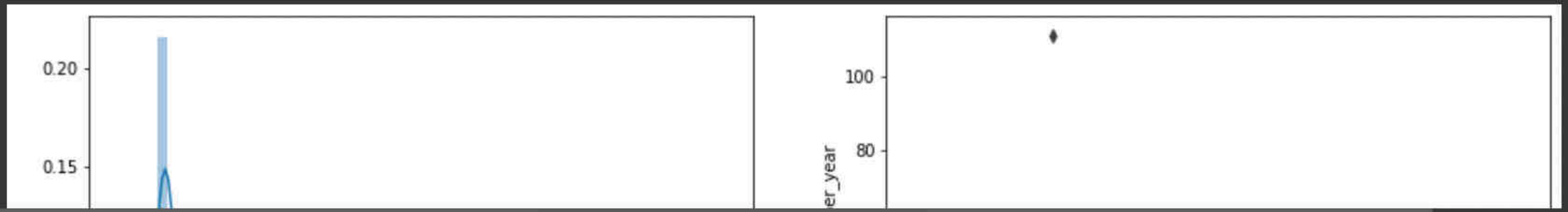


Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:21
10.2 - Hypothesis | Kruskal–Wallis one | Anscombe's quart | Bias of an estimat | EDA_FE.ipynb - C | scipy.stats.krusk | numpy.var — Num | pandas.cut — pan | pandas.qcut — pa | pandas.DataFrame
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=K4uMgmkQNE39 Update
+ Code + Text Reconnect



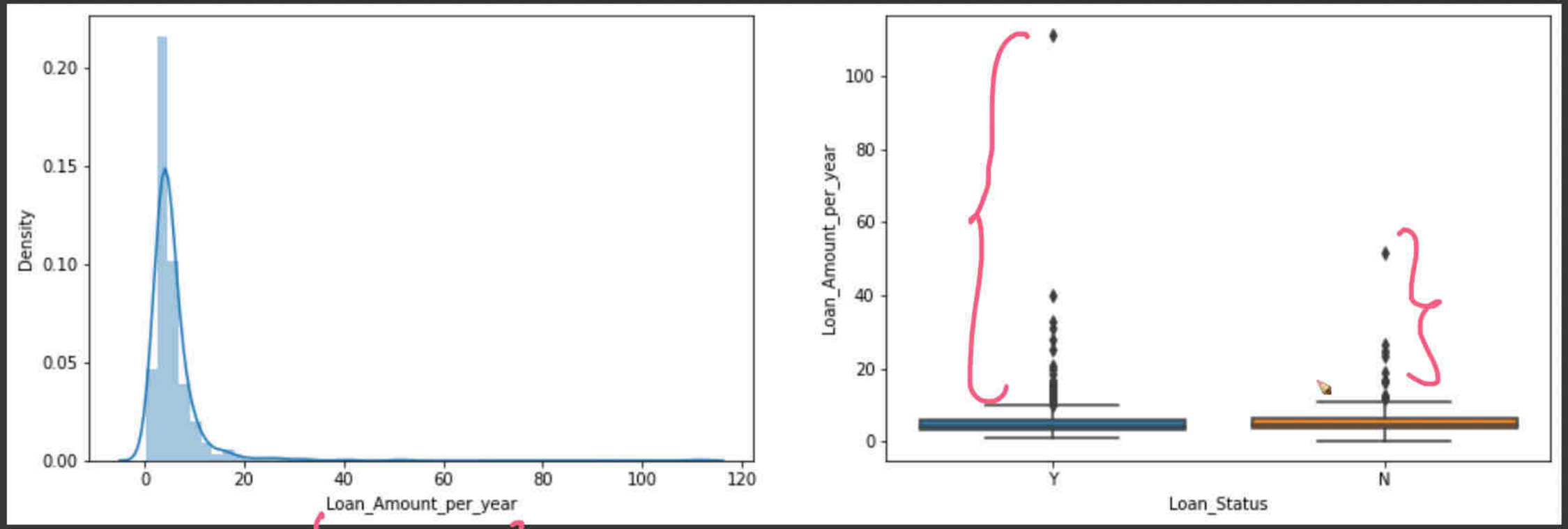

```
# Approximate calc: ignoring interest rates as we dont know that.
data['Loan_Amount_per_year']  = data['LoanAmount']/data['Loan_Amount_Term']
```

```
[ ]  plt.figure(figsize=(16,5))
     plt.subplot(121)
     sns.distplot(data['Loan_Amount_per_year']);

     plt.subplot(122)
     sns.boxplot(data=data, x='Loan_Status', y = 'Loan_Amount_per_year')

     plt.show()
```



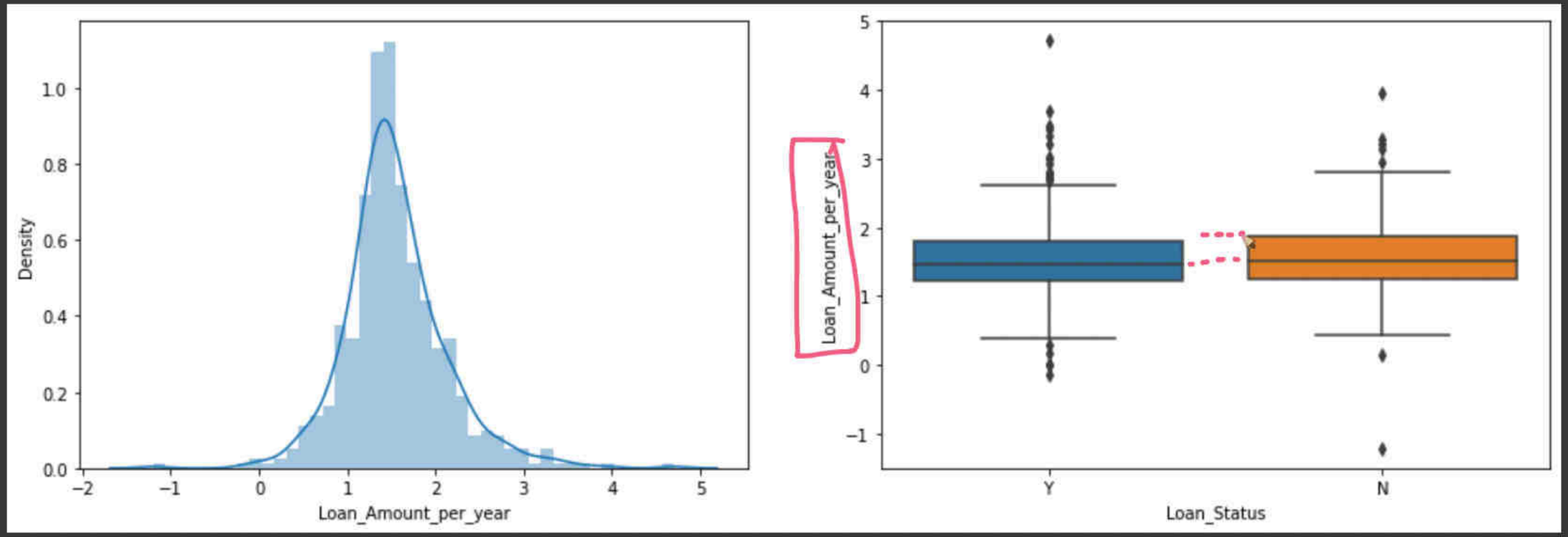
117 / 117

```
sns.boxplot(data=data, x='Loan_Status', y = 'Loan_Amount_per_year')

plt.show()
```

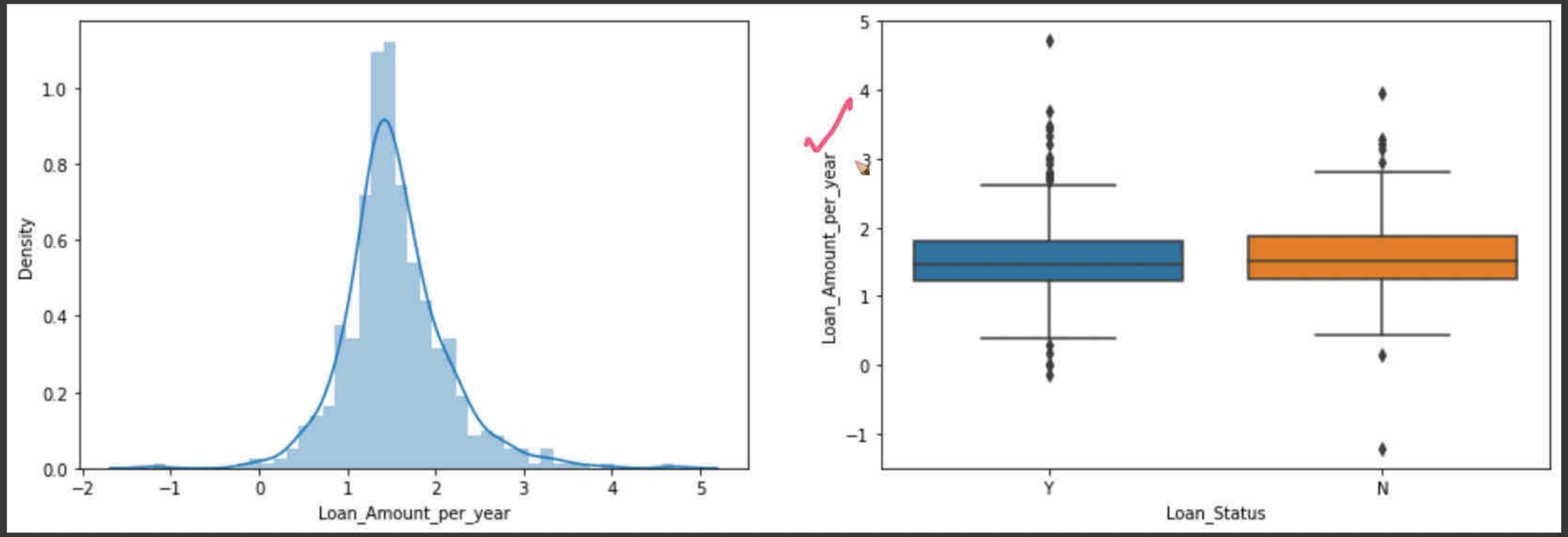



```
[ ] # log transform
    plt.figure(figsize=(16,5))
    plt.subplot(121)
```

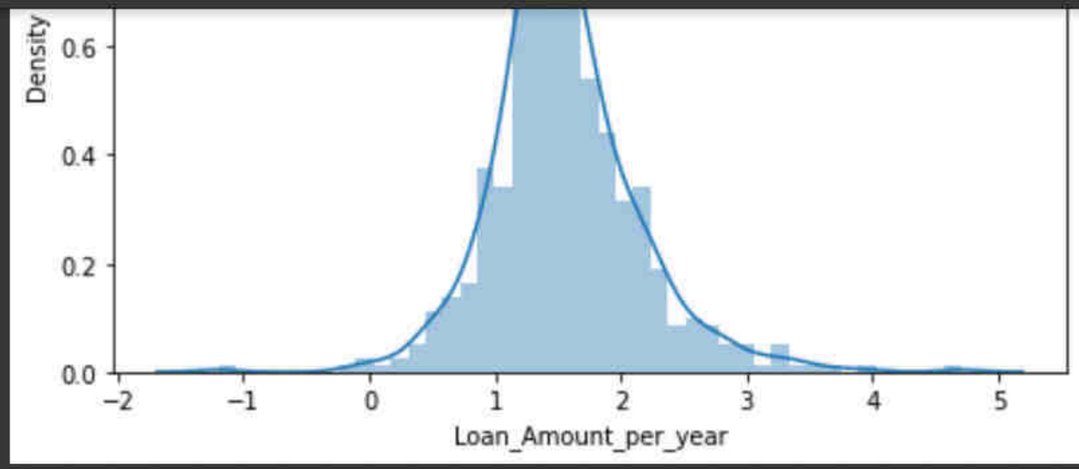
```
sns.distplot(log_loanAmount)

plt.subplot(122)
sns.boxplot(data=data, x='Loan_Status', y = log_loanAmount)

plt.show()
```

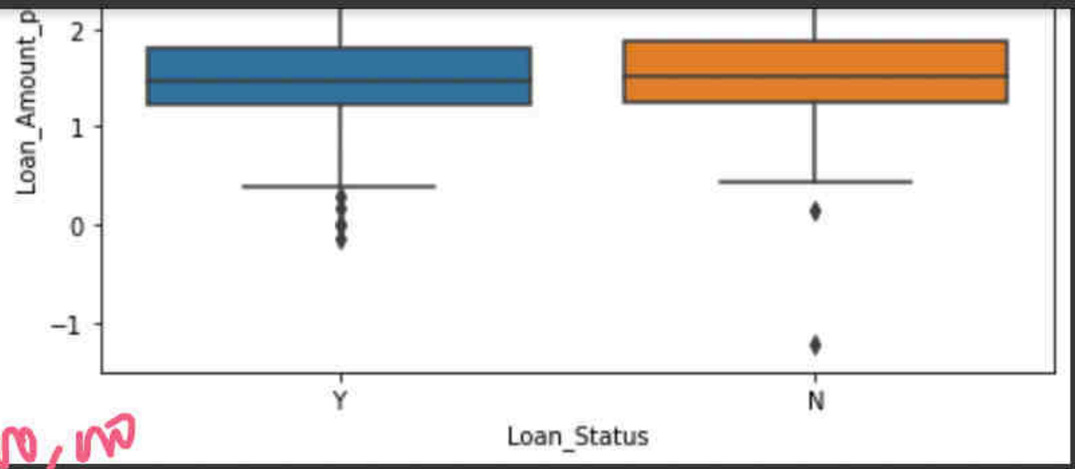


```
sns.distplot(log_loanAmount)

plt.subplot(122)
sns.boxplot(data=data, x='Loan_Status', y = log_loanAmount)

plt.show()
```



Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:23
10.2 - Hypothesis | Kruskal–Wallis one | Anscombe's quart | Bias of an estimat | EDA_FE.ipynb - C | scipy.stats.krusk | numpy.var — Num | pandas.cut — pan | pandas.qcut — pa | pandas.DataFrame
colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=K4uMgmkQNE39 Update
+ Code + Text Reconnect





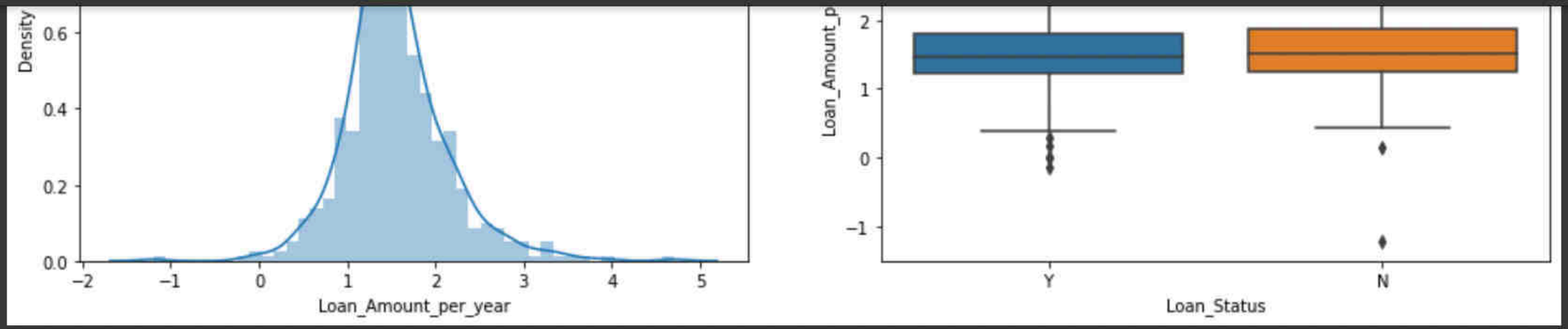
80, 72

```
[ ] # Feature : Calculate the EMI based on the Loan Amount Per year.
    data['EMI'] = data['Loan_Amount_per_year']*1000/12
```

```
[ ] #Feature : Able_to_pay_EMI
    data['Able_to_pay_EMI'] = (data['TotalIncome']*0.1 > data['EMI']).astype('int')
```

```
[ ] sns.countplot(x='Able_to_pay_EMI', data = data, hue = 'Loan_Status')
    #Observation:
    ###There is 50% chance that you may get the loan approved if you cannot pay the EMI.
    ###But there, is a 72% chance that you may get the loan approved if you can pay the EMI.
```

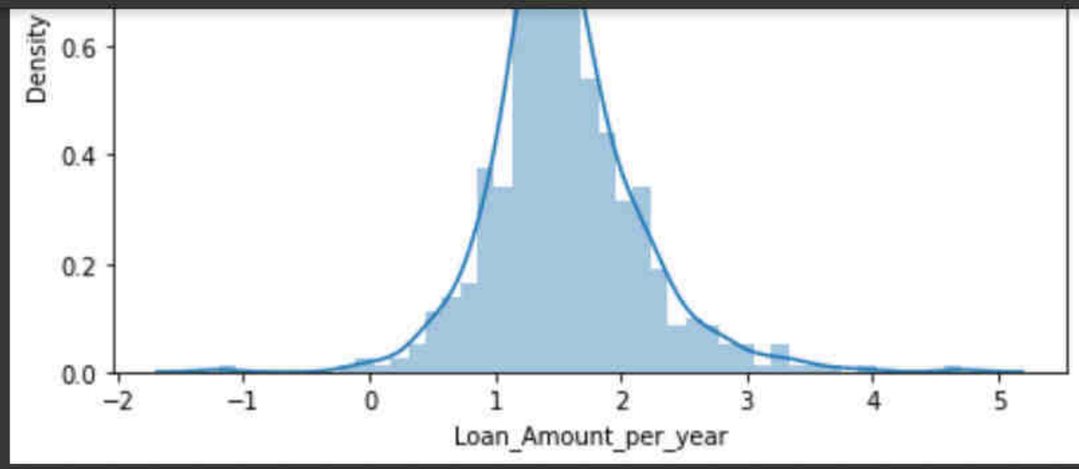
121 / 121

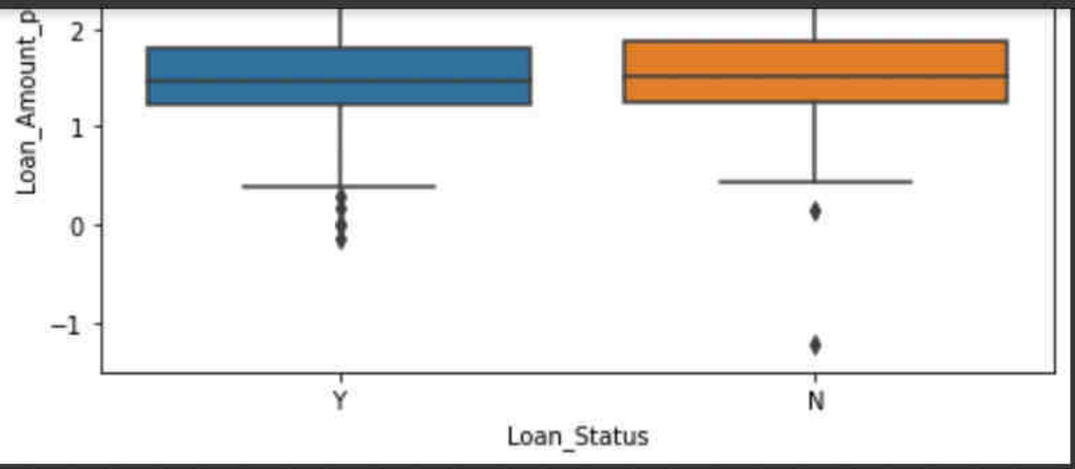



```
[ ] # Feature : Calculate the EMI based on the Loan Amount Per year.
    data['EMI'] = data['Loan_Amount_per_year']*1000/12
```

```
[ ] #Feature : Able_to_pay_EMI
    data['Able_to_pay_EMI'] = (data['TotalIncome']*0.1 > data['EMI']).astype('int')
```

```
[ ] sns.countplot(x='Able_to_pay_EMI', data = data, hue = 'Loan_Status')
    #Observation:
    ###There is 50% chance that you may get the loan approved if you cannot pay the EMI.
    ###But there, is a 72% chance that you may get the loan approved if you can pay the EMI.
```






```
[ ] # Feature : Calculate the EMI based on the Loan Amount Per year.
    data['EMI'] = data['Loan_Amount_per_year']*1000/12
```

emi = loan-amt x 1000 / duration / 12

```
[ ] #Feature : Able_to_pay_EMI
    data['Able_to_pay_EMI'] = (data['TotalIncome']*0.1 > data['EMI']).astype('int')
```

```
[ ] sns.countplot(x='Able_to_pay_EMI', data = data, hue = 'Loan_Status')
    #Observation:
    ###There is 50% chance that you may get the loan approved if you cannot pay the EMI.
    ###But there, is a 72% chance that you may get the loan approved if you can pay the EMI.
```

```
#Feature : Able_to_pay_EMI
data['Able_to_pay_EMI'] = (data['TotalIncome']*0.1 > data['EMI']).astype('int')
```

```
[ ] sns.countplot(x='Able_to_pay_EMI', data = data, hue = 'Loan_Status')
    #Observation:
    ###There is 50% chance that you may get the loan approved if you cannot pay the EMI.
    ###But there is a 72% chance that you may get the loan approved if you can pay the EMI.
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3e2eadc950>
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:38

10.2 - Hypothe | Kruskal–Wallis | Anscombe's qu | Bias of an estim | EDA_FE.ipynb | scipy.stats.krus | numpy.var — N | pandas.cut — p | pandas.qcut — | pandas.DataFra | Pareto distribut

colab.research.google.com/drive/1_P-nKswaKnx5v77lOx3pCxvJve73MvnN#scrollTo=FxqCjW7IaETR Update

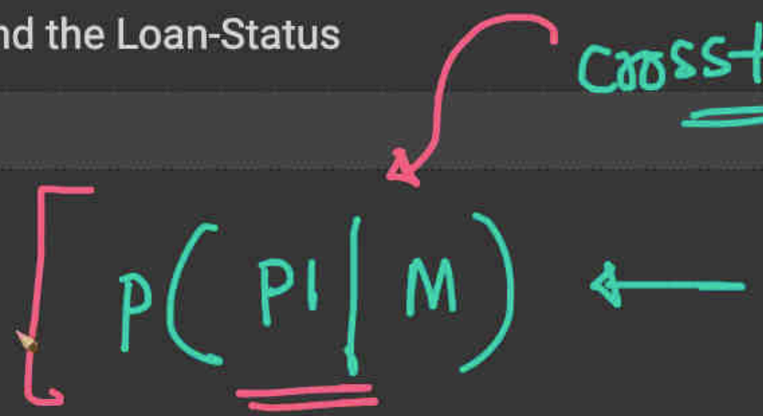
+ Code + Text Reconnect

3. Target Encoding

Appropriate encoding depends on what our task is (and) what we do next?

Task: Compute Correlation (PCC and SRCC) between each feature and the Loan-Status

[ ] ↳ *30 cells hidden*



Column Standarization and Normalization

- Mean centering and Variance scaling (Stndard Scaling)
- MinMax Scaling

[ ] ↳ *1 cell hidden*

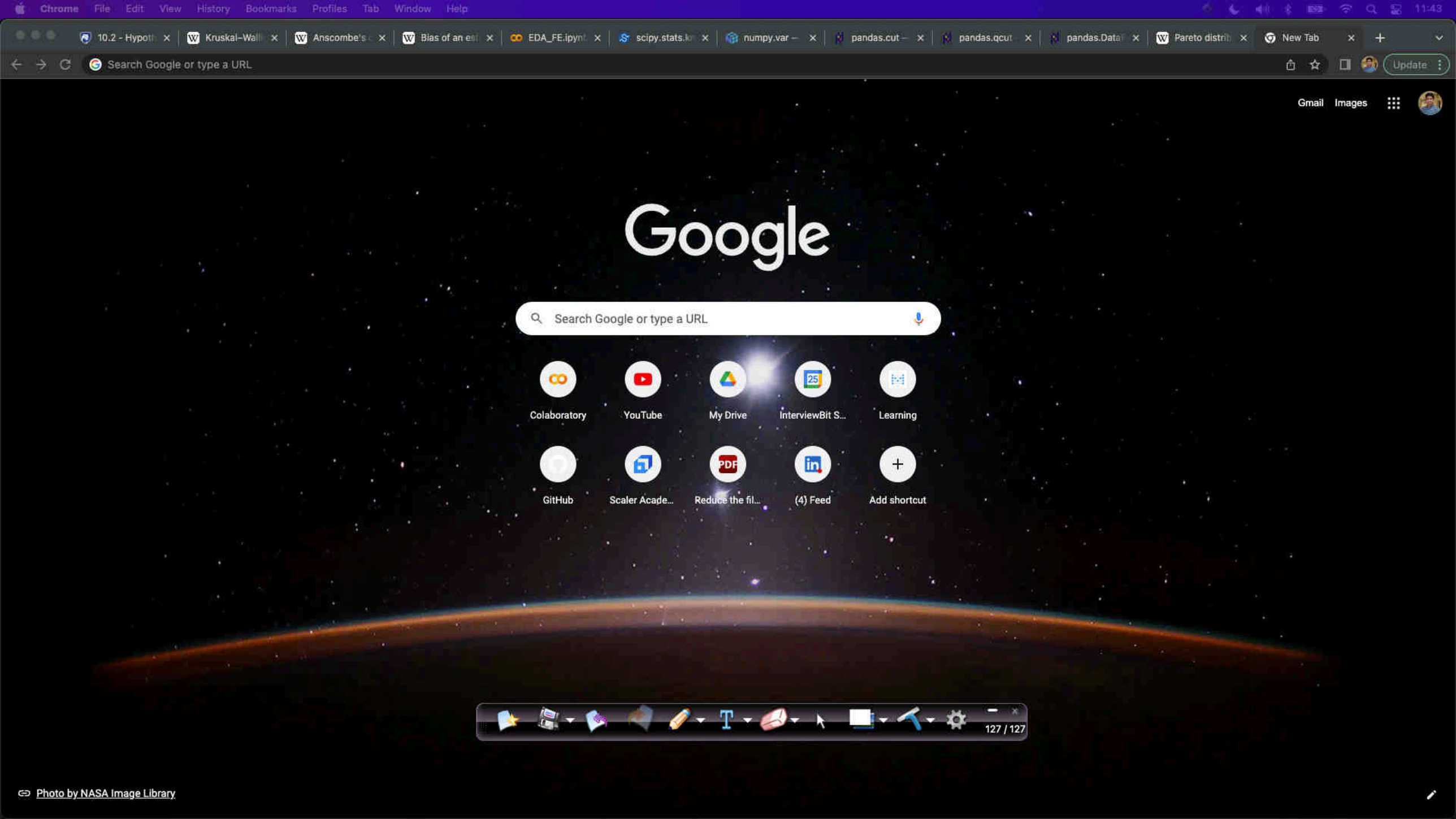
125 / 125

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:42

Q

$$T_{\chi^2} = \sum_{i=1}^{\#cells} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k)$$

$k = (n_{rows} - 1) * (n_{cols} - 1)$


Chrome File Edit View History Bookmarks Profiles Tab Window Help
10.2 - Hypoth
Kruskal–Wall
Anscombe's
Bias of an est
EDA_FE.ipynb
scipy.stats.lo
numpy.var –
pandas.cut –
pandas.qcut
pandas.Data
Pareto distrib
New Tab
Search Google or type a URL
Update
Gmail
Images
Google
Search Google or type a URL
Colaboratory
YouTube
My Drive
InterviewBit S...
Learning
GitHub
Scaler Acade...
Reduce the fil...
(4) Feed
Add shortcut
127 / 127
Photo by NASA Image Library