STATISTICS PROJECT

Presentation by Ingrid Diniz, Igor Diniz and José Santos

Master in Engineering and Data Science | 2023

FEUP

PROJECT DESCRIPTION

STATISTICAL ANALYSIS INVOLVING:

Profile of FAB Aircraft Involved in Accidents and Possible Correlations Between Aircraft Age, Contributing Factors, Occurrence Type, Weather, and Fatalities

Presentation by Ingrid Diniz, Igor Diniz and José Santos

DATA DESCRIPTION

The aeronautical occurrences database is maintained by the Aeronautical Accident Investigation and Prevention Center (CENIPA). This database covers the period from 2010 to 2021 and comprises the record of aeronautical occurrences reported to CENIPA that occurred in Brazilian territory.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

METHODOLOGY

SIX-STEP STATISTICAL INVESTIGATION METHOD

1. Ask a reserarch question
2. Design a study and collect data
3. Explore the data
4. Draw inferences
5. Formulate conclusions
6. Look back and ahead

Presentation by Ingrid Diniz, Igor Diniz and José Santos

RESERACH QUESTION

IS THE AGE OF AN AIRCRAFT RELATED TO OCCURRENCES?

Presentation by Ingrid Diniz, Igor Diniz and José Santos

STUDY AND DATA

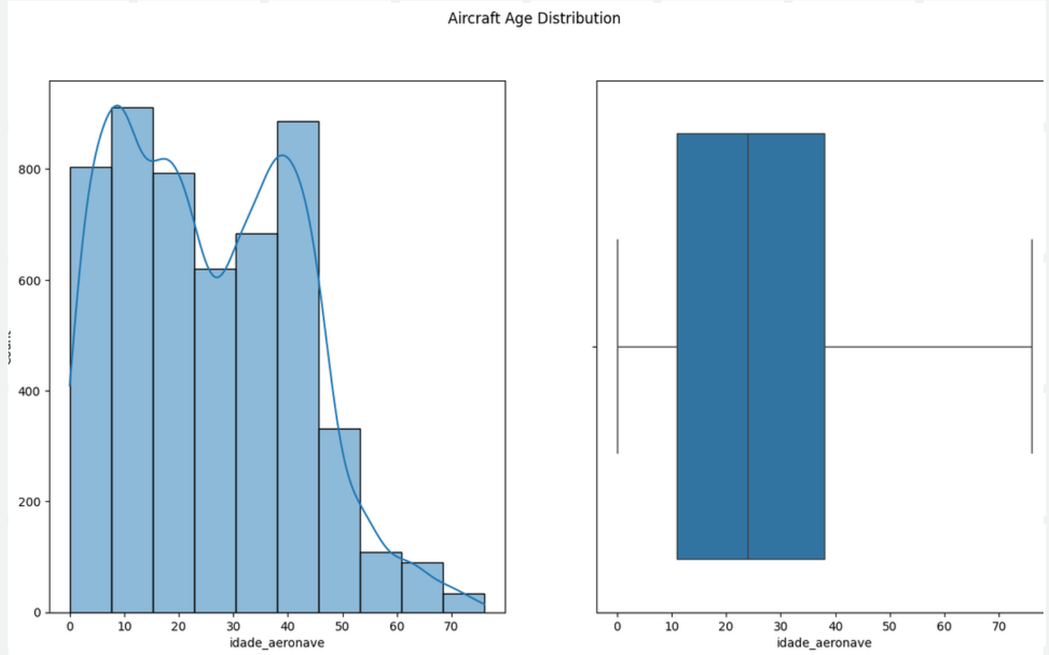
For the search, we collect the data from FAB. More specifically, we worked with four different csv files containing data from "ocorrencias", "tipo de ocorrencia", "aeronave" e "fator".
We also went to the National Institute of Meteorology (INMET) to retrieve weather data so that we could analyze how the wind may impact aircraft occurrences.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

EXPLORATORY DATA ANALYSIS

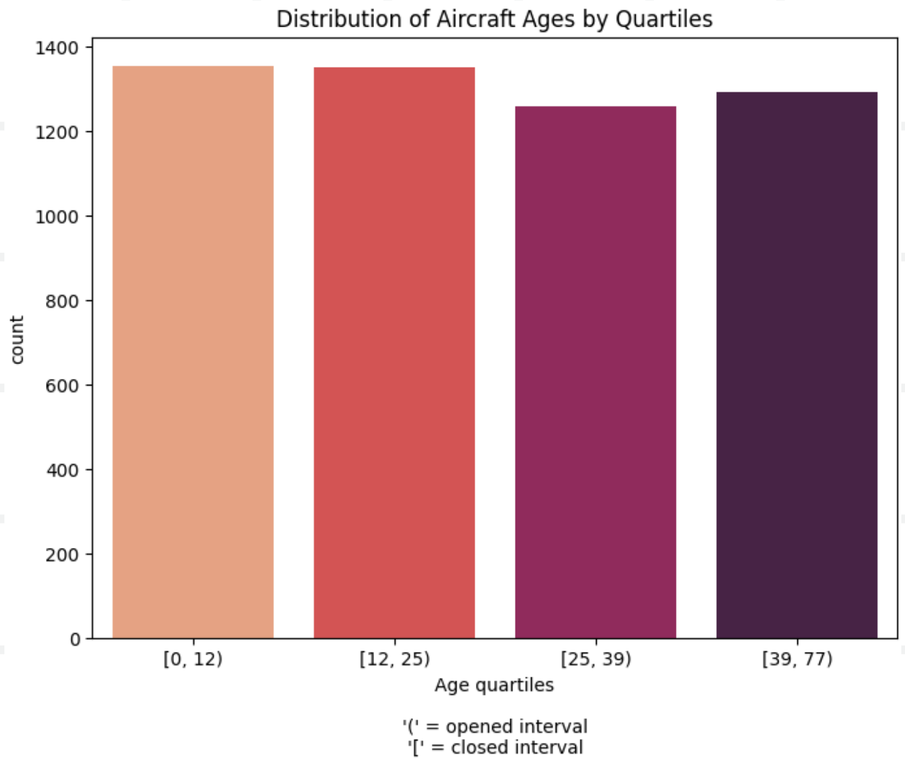
Age Distribution of Aircrafts

	aeronave_ano_fabricacao	aeronave_fatalidades_total	idade_aeronave
count	5258.000000	5258.000000	5258.000000
mean	1990.858692	0.629897	25.534994
std	16.222558	1.454515	16.230823
min	1945.000000	0.000000	0.000000
25%	1977.000000	0.000000	11.000000
50%	1992.000000	0.000000	24.000000
75%	2006.000000	0.000000	38.000000
max	2021.000000	10.000000	76.000000



Presentation by Ingrid Diniz, Igor Diniz and José Santos

EXPLANING WHAT IS CONSIDERED NEW AND OLD AIRCRAFTS

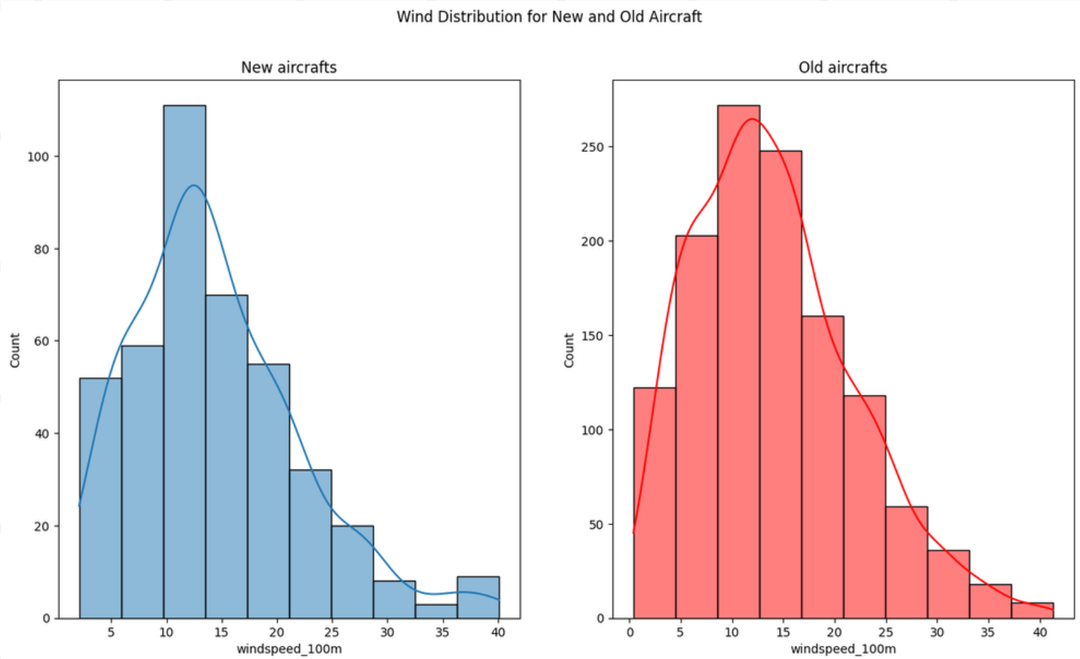


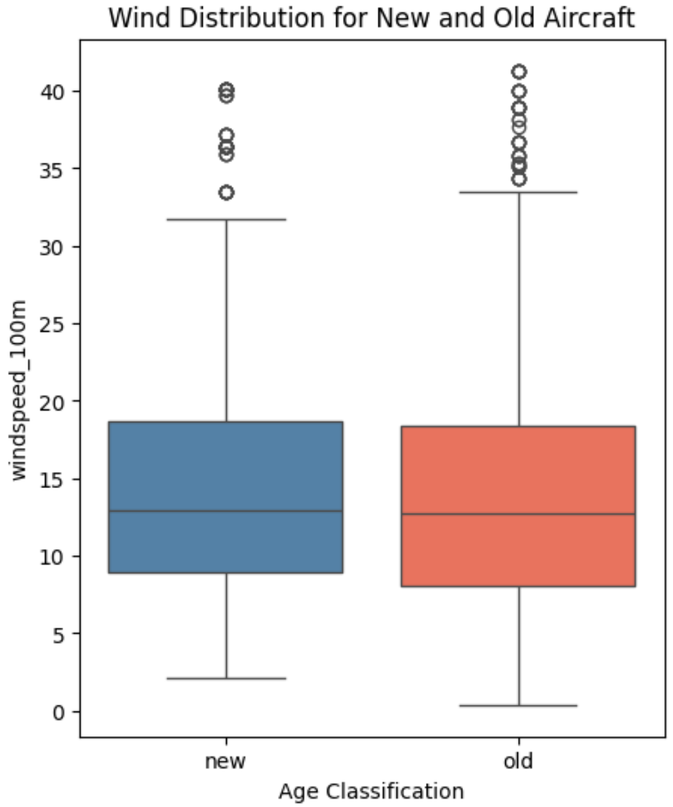
- We discretized aircraft ages into quartiles to distinguish between new and old airplanes.
- After obtaining [11, 24, 38] for the [25%, 50%, 75%] quantiles respectively, we decided that new airplanes are those with an age less than or equal to 11, which corresponds to the first quartile (25%). Conversely, old aircraft are those with an age greater than 11.
- We chose to use quartiles because it allows for a roughly equal distribution of occurrences in each discretized interval, as illustrated in the chart.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

WIND DISTRIBUTION

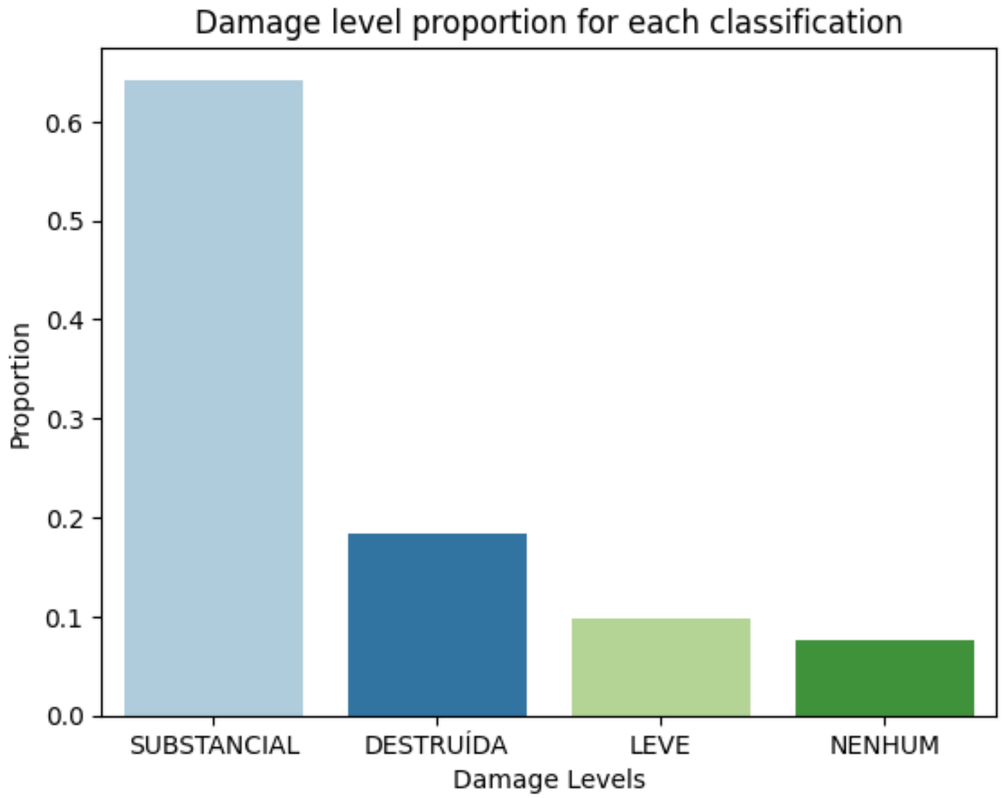
Average wind speed for occurrences

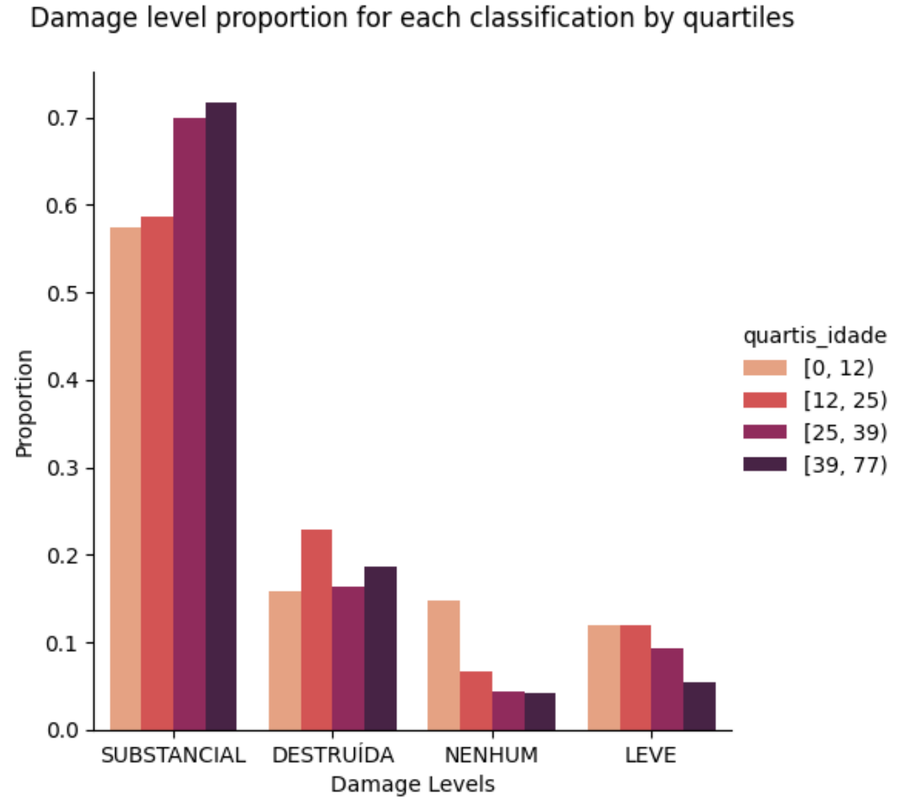




Presentation by Ingrid Diniz, Igor Diniz and José Santos

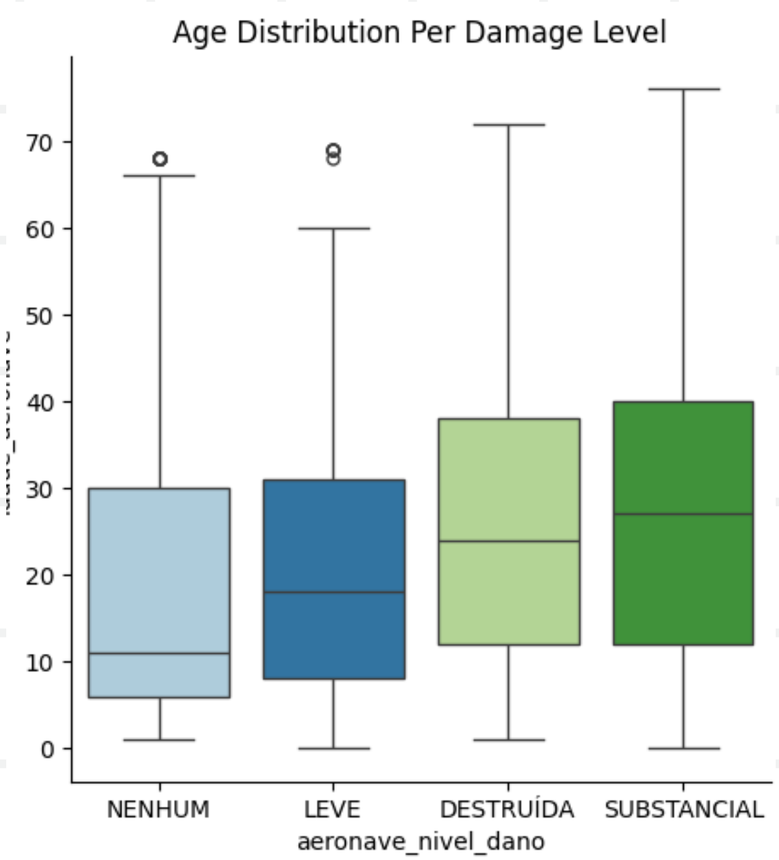
DAMAGE LEVEL ANALYSIS

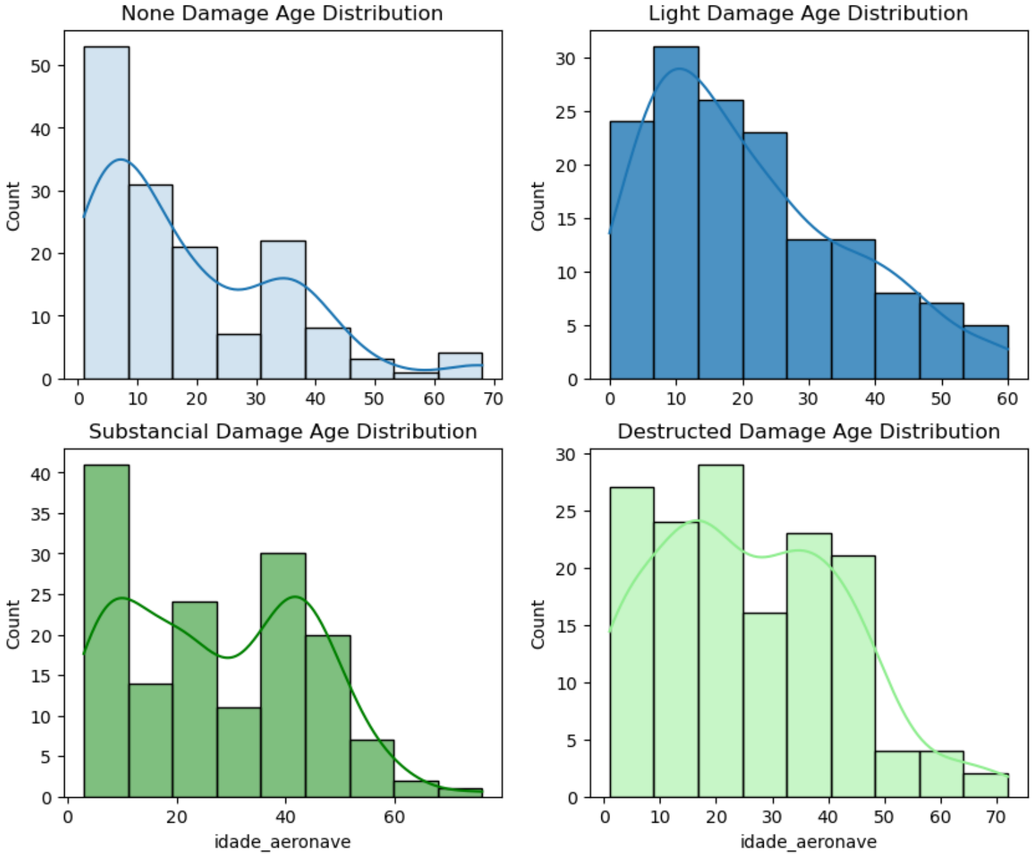




Presentation by Ingrid Diniz, Igor Diniz and José Santos

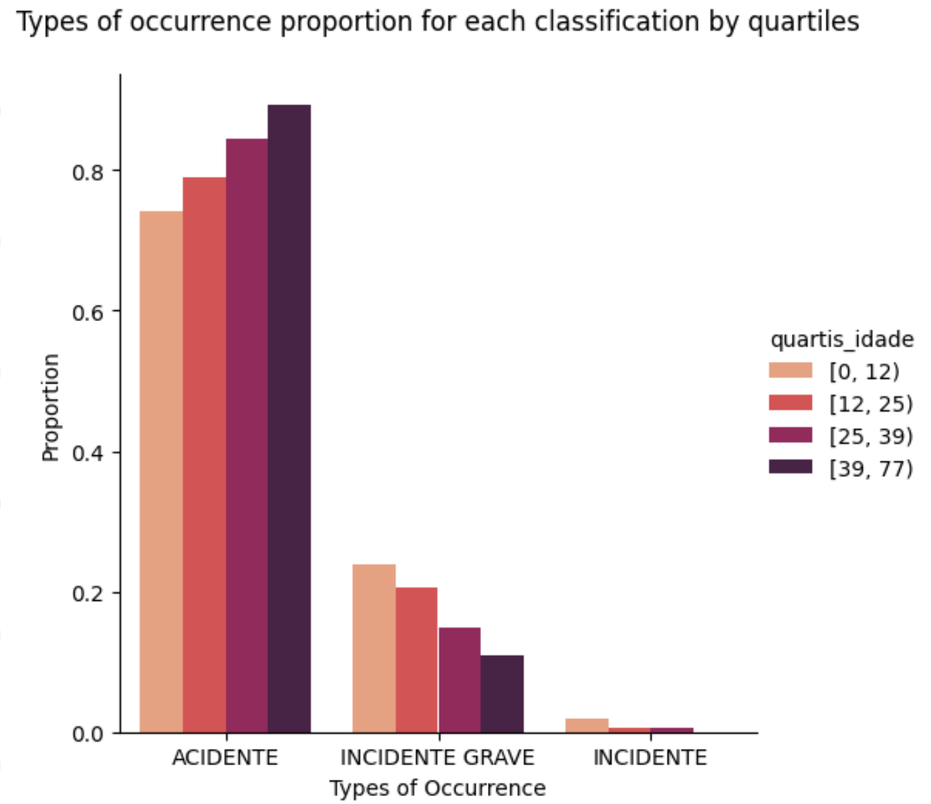
DAMAGE LEVEL ANALYSIS

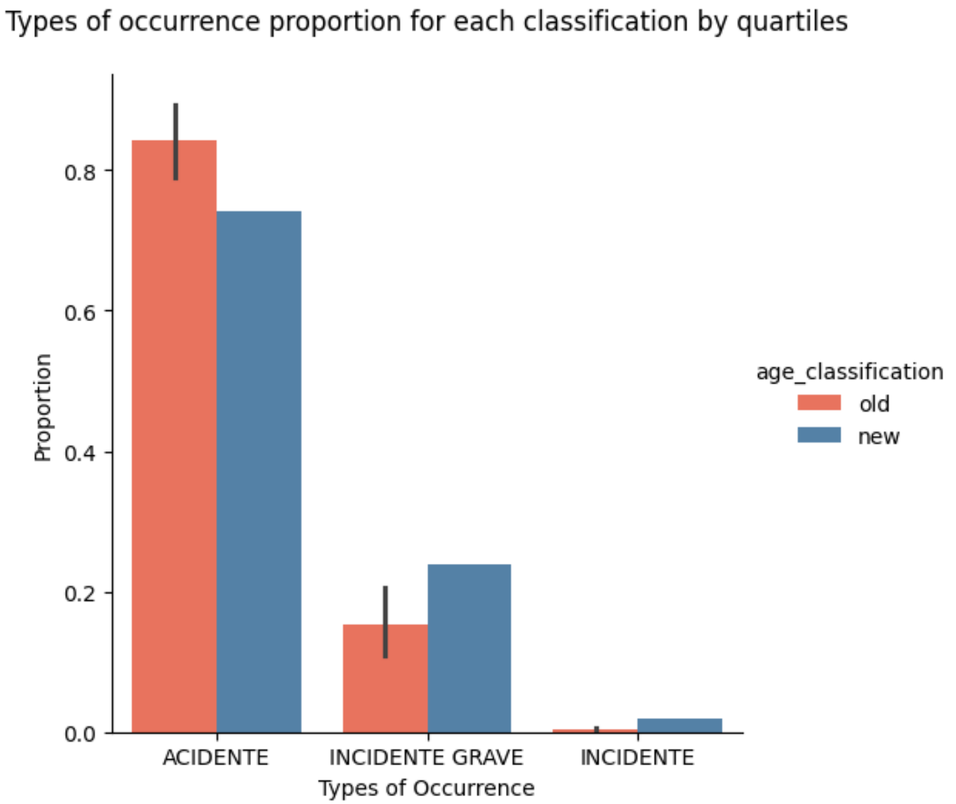




Presentation by Ingrid Diniz, Igor Diniz and José Santos

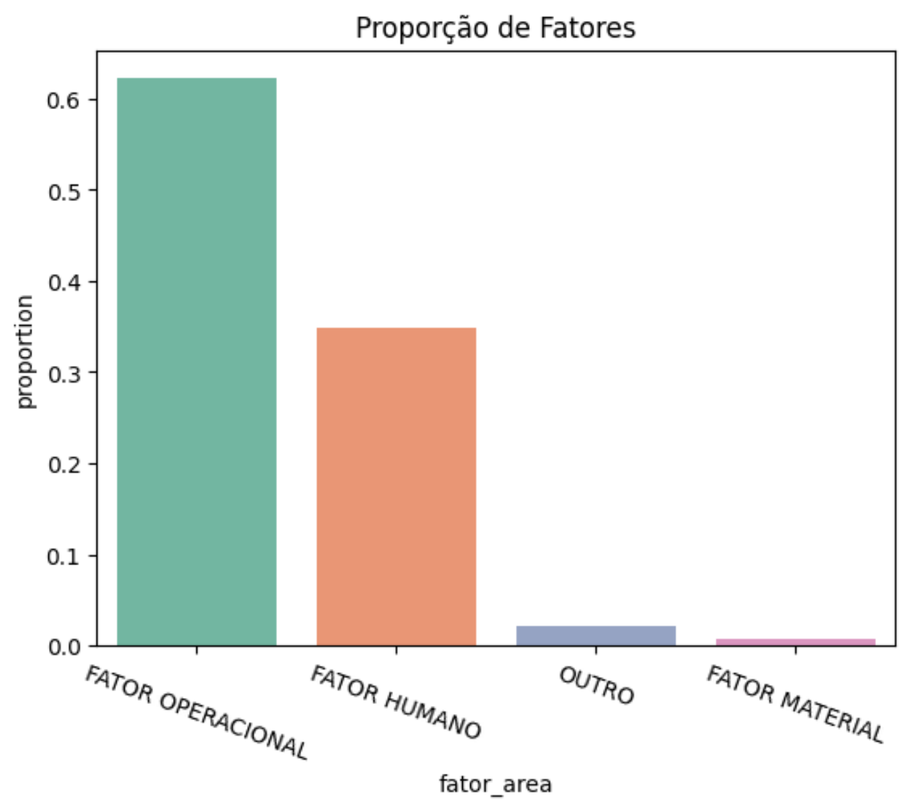
TYPE OF OCCURRENCE ANALYSIS

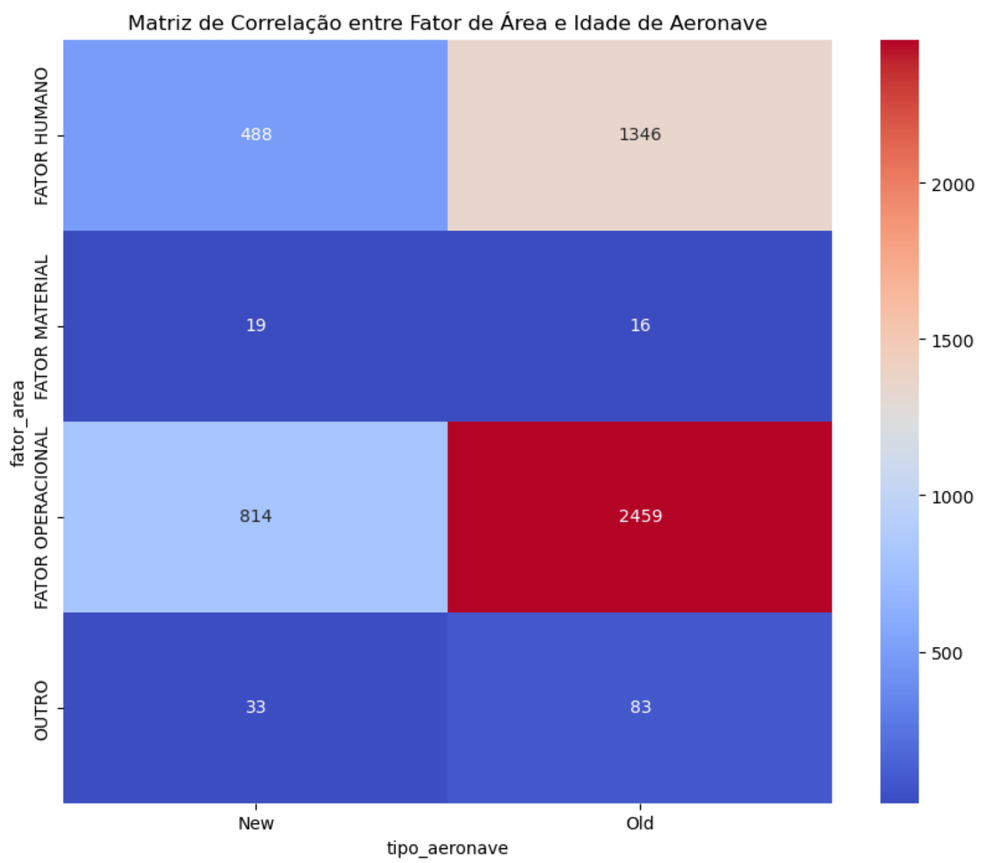




Presentation by Ingrid Diniz, Igor Diniz and José Santos

FACTOR ANALYSIS





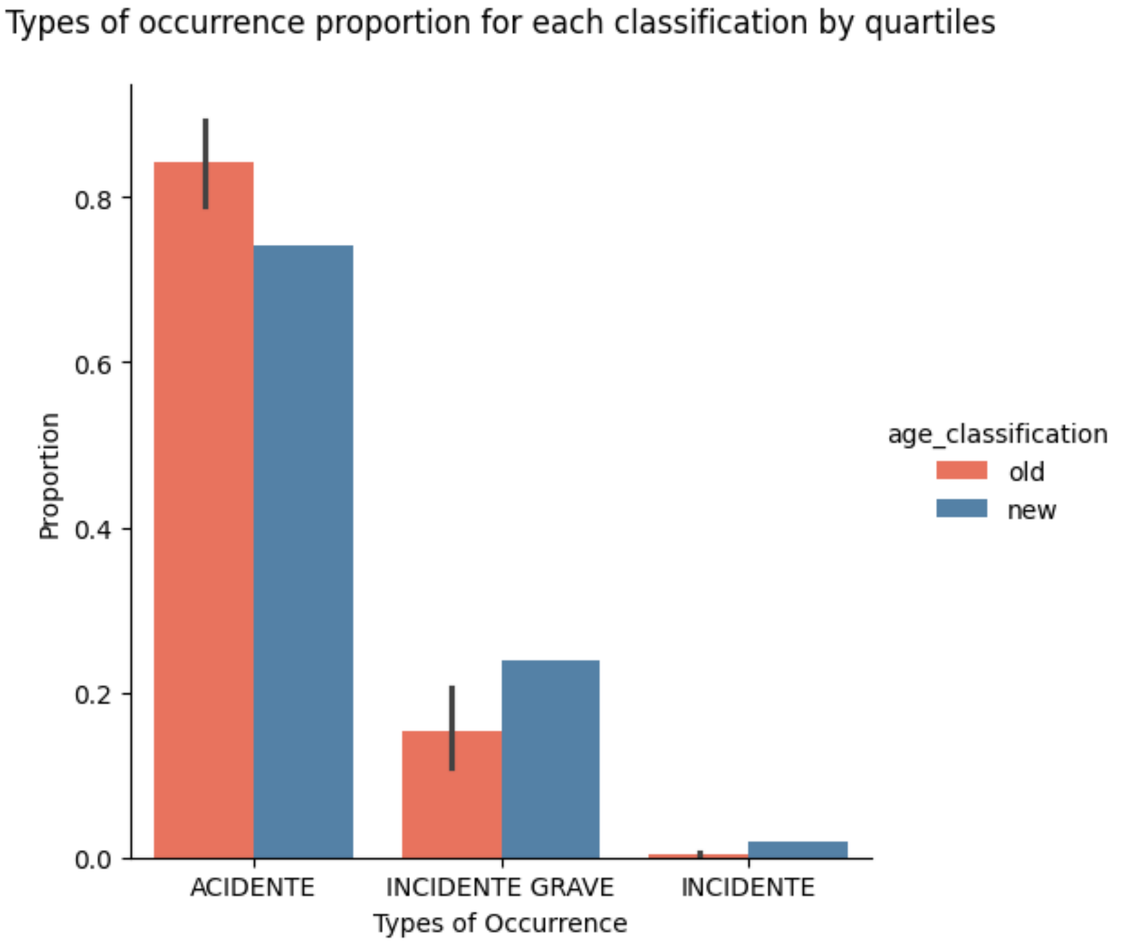
Presentation by Ingrid Diniz, Igor Diniz and José Santos

DRAWING INFERENCES

- Is there a difference between the proportion of accidents in older aircraft and newer aircrafts?
- Are there differences among the averages of airplanes in each damage level? If so, which ones?
- Is there a significant relationship between the occurrence factor and the age of the aircraft?
- Are older aircraft more susceptible to accidents due to wind?

Presentation by Ingrid Diniz, Igor Diniz and José Santos

IS THERE A DIFFERENCE BETWEEN THE PROPORTION OF ACCIDENTS IN OLDER AIRCRAFT AND NEWER AIRCRAFT?



Hypothesis Formulation

- H_0: $p>11 - p\leq11 = 0$
- Ha: $p>11 - p\leq11 \neq 0$

Presentation by Ingrid Diniz, Igor Diniz and José Santos

IS THERE A DIFFERENCE BETWEEN THE PROPORTION OF ACCIDENTS IN OLDER AIRCRAFT AND NEWER AIRCRAFT?

Checking conditions

- **Independence:** Because the data come from a simple random sample, the observations are independent, both within and between samples.

- **Success-failure condition:** Since the null hypothesis is that the proportions are equal, we use the pooled proportion (p̂pooled) to verify the success-failure condition and estimate the standard error.

$$\hat{p} * n \geq 10 \text{ and } (\hat{p} - 1) * n \geq 10$$

- For n1 = n2 = 150, p̂≤11 = 0.75, p̂>11 = 0.9 and p̂pooled = 0.83 both conditions above are met.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

IS THERE A DIFFERENCE BETWEEN THE PROPORTION OF ACCIDENTS IN OLDER AIRCRAFT AND NEWER AIRCRAFT?

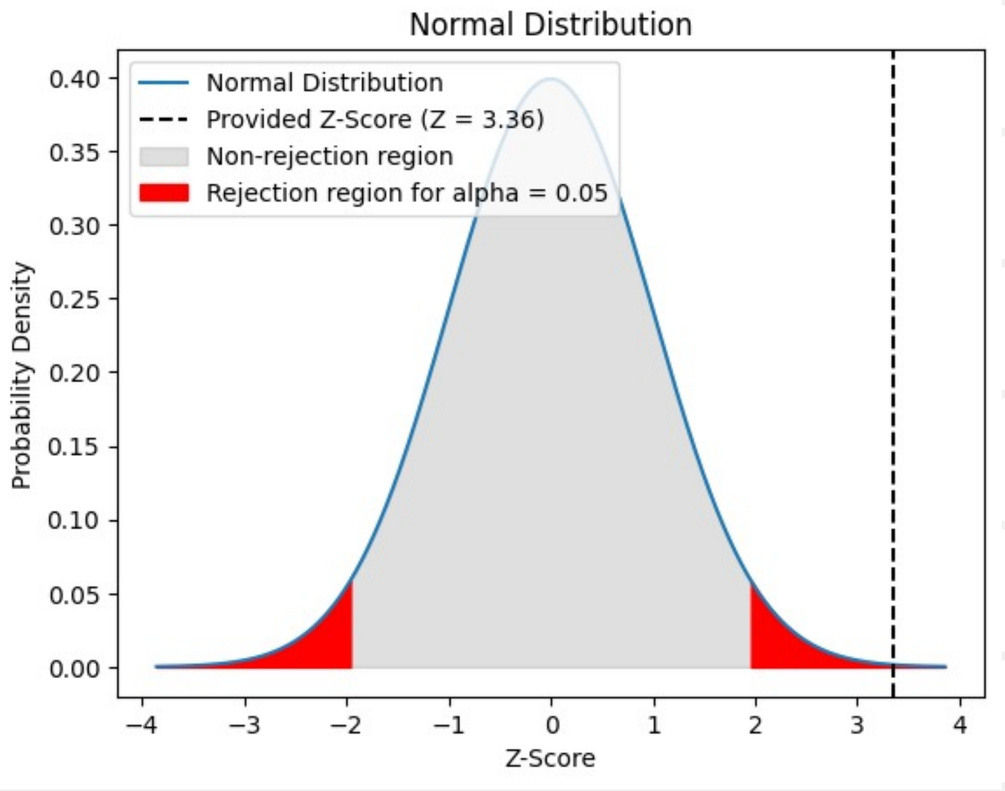
Standard error for the difference of proportions

$$SE = \sqrt{\frac{\hat{p}_{pooled} * (1-\hat{p}_{pooled})}{n_1} + \frac{\hat{p}_{pooled} * (1-\hat{p}_{pooled})}{n_2}}$$

Test statistic

$$Z^* = \frac{point\ estimate - null\ value}{SE}$$

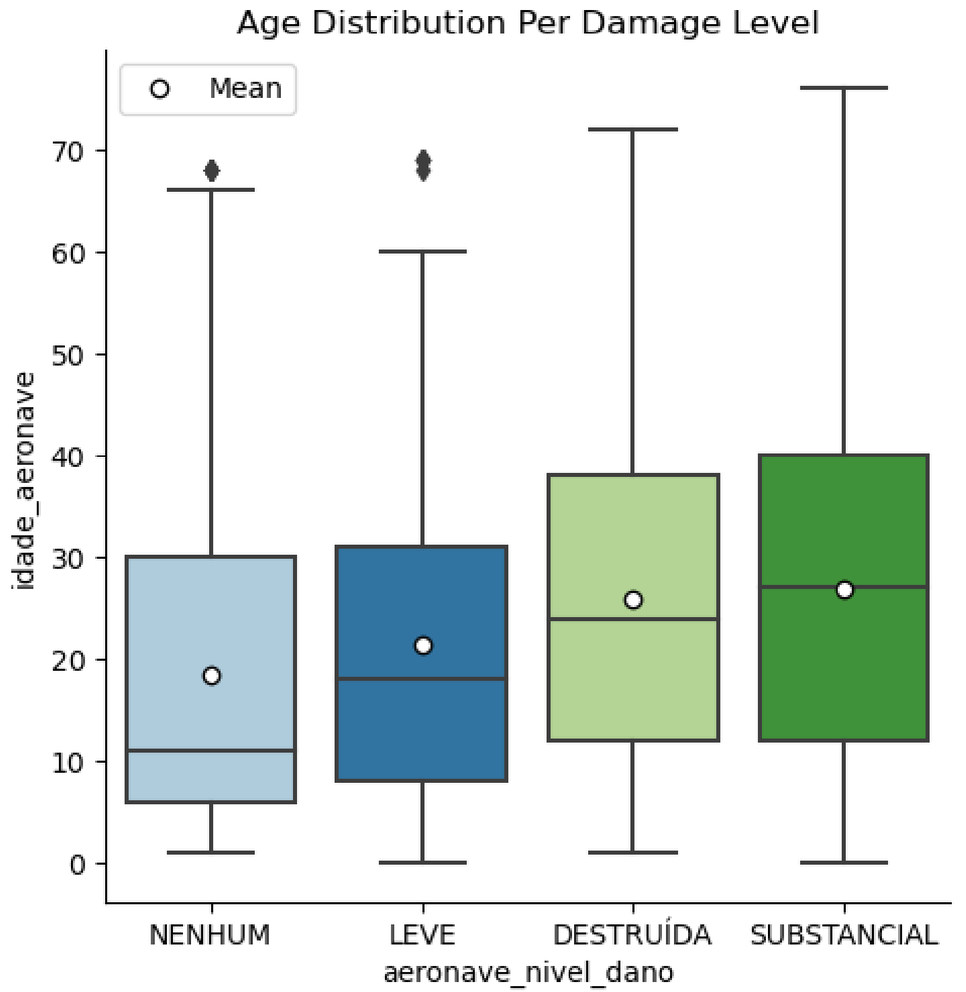
- point estimate = $\hat{p}_{>11} - \hat{p}_{\leq 11} = 0.146$
- SE = 0.044
- Z*= 3.35
- Confidence Level = 0.95
- p_value = 0.00079
- Confidence Interval = (0.061, 0.232)



RESULT: WE REJECTED THE NULL HYPOTHESIS!

Presentation by Ingrid Diniz, Igor Diniz and José Santos

ARE THERE DIFFERENCES AMONG THE AVERAGES OF AGE OFAIRPLANES IN EACH DAMAGE LEVEL? IF SO, WHICH ONES?

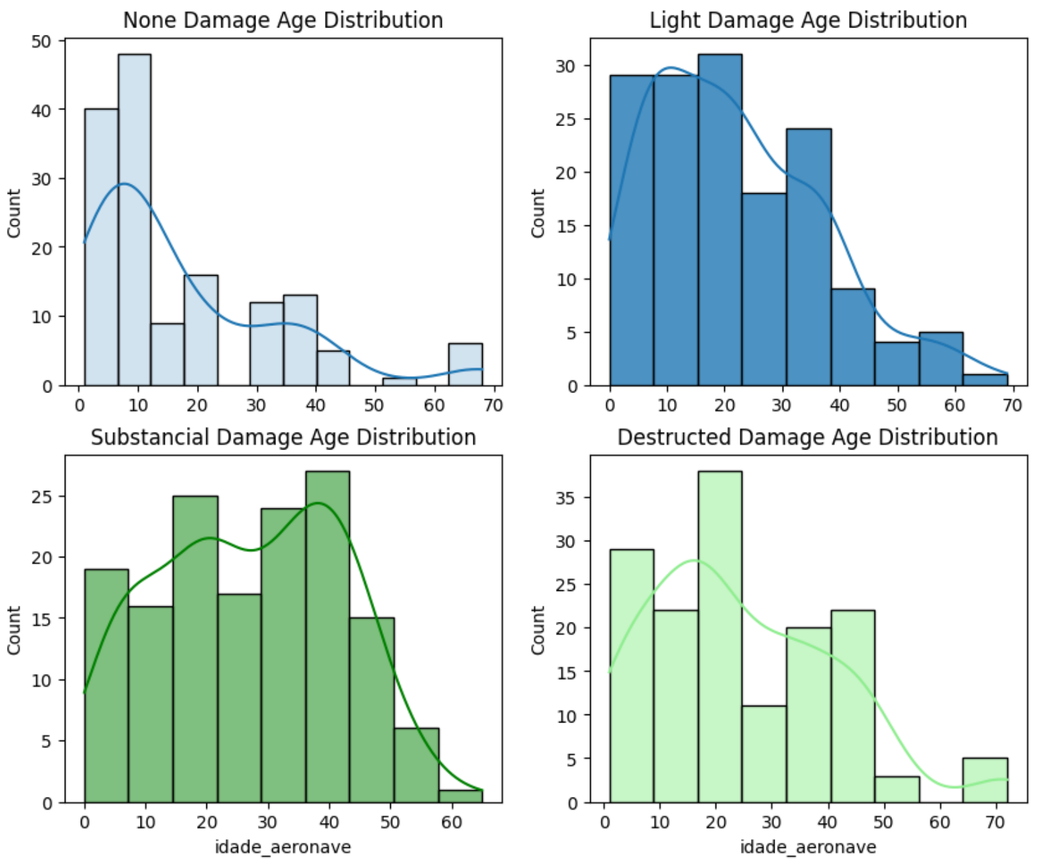


Hypothesis Formulation

- **H_0: $\mu_{NENHUM} = \mu_{SUBSTANCIAL} = \mu_{LEVE} = \mu_{DESTRUIDA}$**
- **Ha:** The average of airplane ages **(μ)** varies across some (or all) groups.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

ARE THERE DIFFERENCES AMONG THE AVERAGE OF AGE OF AIRPLANES IN EACH DAMAGE LEVEL? IF SO, WHICH ONES?



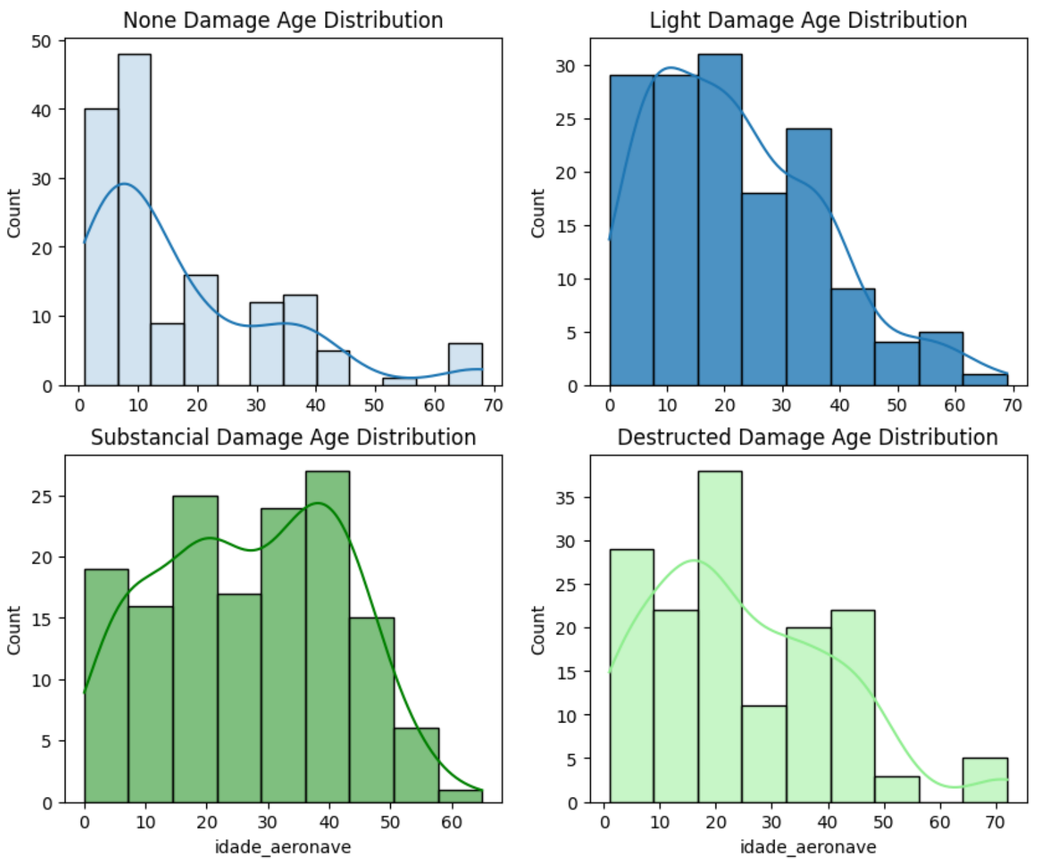
Checking conditions

- **Independence between and within each category**: Since we are taking simple random samples for each category with **n=250** this condition is checked.

- **Approximately normal distribution of each group:** We can check by looking at the charts that the normality is not so approximated, even more in the None damaged one, which is a right-skewed chart. However, since we are taking large samples (n=250) we will relax this point.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

ARE THERE DIFFERENCES AMONG THE AVERAGE OF AGE OF AIRPLANES IN EACH DAMAGE LEVEL? IF SO, WHICH ONES?



Checking conditions

- **Independence between and within each category**: Since we are taking simple random samples for each category with **n=250** this condition is checked.

- **Approximately normal distribution of each group:** We can check by looking at the charts that the normality is not so approximated, even more in the None damaged one, which is a right-skewed chart. However, since we are taking large samples (n=250) we will relax this point.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

ARE THERE DIFFERENCES AMONG THE AVERAGES OF AGE OF AIRPLANES IN EACH DAMAGE LEVEL? IF SO, WHICH ONES?

Checking conditions

- **Constant variance comparing the groups**: We can consider the variances for each category approximately equal, for these taken samples. So, this condition is also checked.
- Age variance in the None Damage sample: 275.98
- Age variance in the Light Damage sample: 258.46
- Age variance in the Substantial Damage sample: 271.03
- Age variance in the Destructed Damage sample: 294.67

- LeveneResult:
  - statistic=2.523
  - pvalue=0.0565

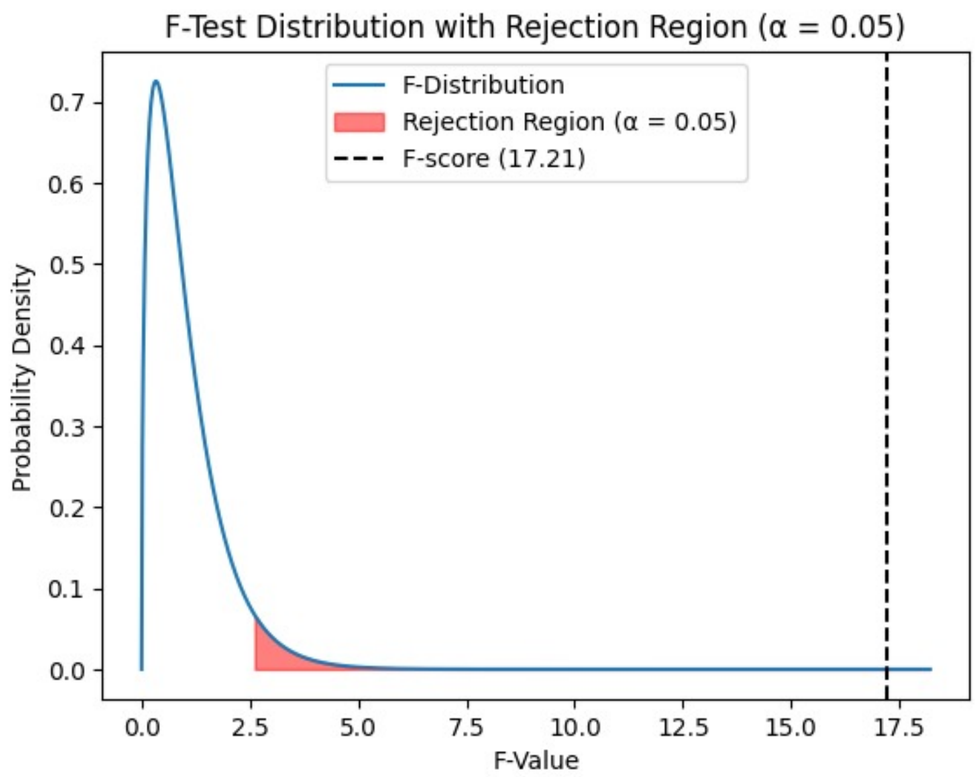
→ The test is not significant (p-value > 0.05), meaning that there is homogeneity of variances and we can proceed with ANOVA.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

ARE THERE DIFFERENCES AMONG THE AVERAGES OF AGE OF AIRPLANES IN EACH DAMAGE LEVEL? IF SO, WHICH ONES?

- We applied ANOVA to check if there were differences between the average ages across the 4 groups.

- n_{NENHUM} $n_{SUBSTANCIAL}$ n_{LEVE} $n_{DESTRUIDA}$ = 250
- $\alpha = 0.05$
- dfG = 3
- dfE = 996
- F-score : 17.21
- p-value ~ 0



RESULT: WE REJECTED THE NULL HYPOTHESIS!

Presentation by Ingrid Diniz, Igor Diniz and José Santos

T-TEST FOR EACH PAIR WITH BONFERRONI CORRECTION FOR ALPHA

We still do not know which levels are different from others. To do that, we will apply the t-test with some modifications for each pair of categories and check which pairs reject the null hypothesis.

Applying BONFERRONI correction for α

$$\alpha^*_{BONFERRONI} = \frac{\alpha}{number\ of\ pairs\ to\ compare}$$

Comparing each pair

Standard error for the difference of proportions

$$SE = \sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}}$$

Spooled = 16.49

Test statistic

$$T^* = \frac{point\ estimate - null\ value}{SE}$$

point estimate = $\bar{x}1 - \bar{x}2$

Significance level we will use

$$\alpha * (Bonferroni) = 0.0083$$

Presentation by Ingrid Diniz, Igor Diniz and José Santos

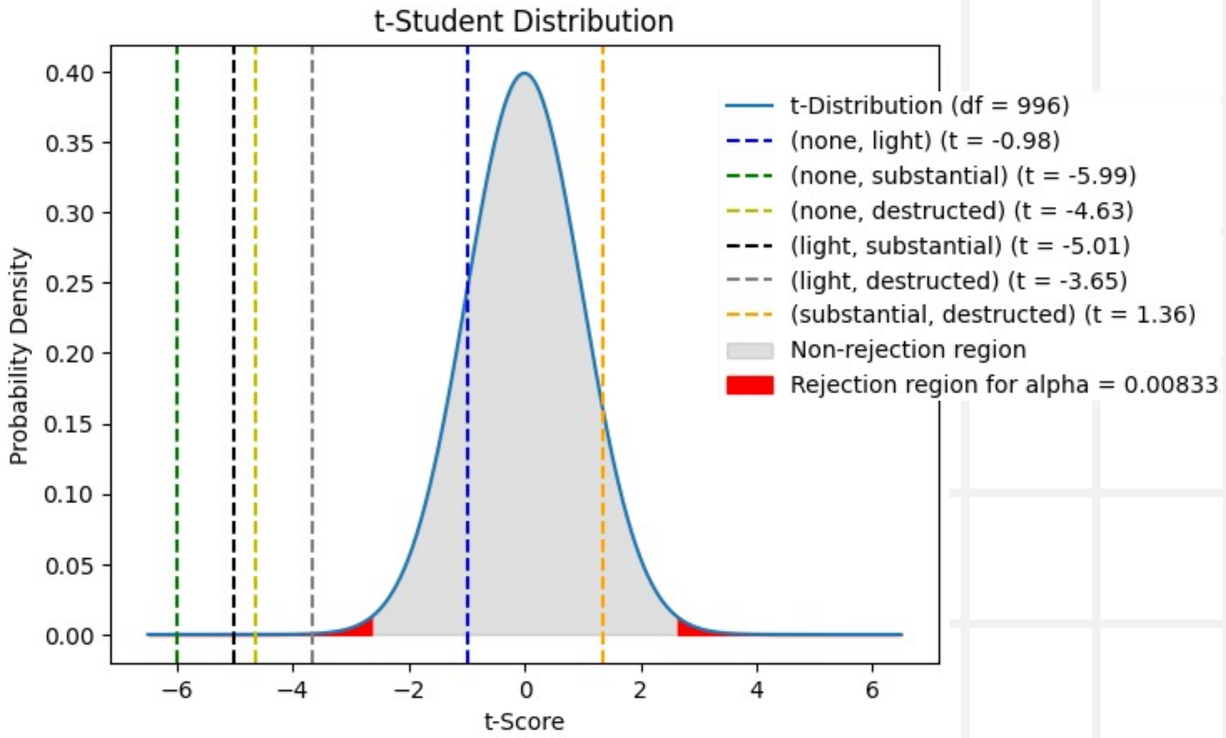
HYPOTHESIS FORMULATION FOR APPLYING T-TEST FOR EACH PAIR OF DAMAGE LEVEL

For each pair being evaluated using the t-test with the BONFERRONI correction we are stating the following hypothesis:

Hypothesis Formulation

- **$H_0: \mu_x = \mu_y$**
- **$Ha: \mu_x \neq \mu_y$**

Where, x and y belongs to {NONE, LIGHT, SUBSTANTIAL, DAMAGE} and x ≠ y



Presentation by Ingrid Diniz, Igor Diniz and José Santos

T-TEST FOR EACH PAIR WITH BONFERRONI CORRECTION FOR ALPHA

Results for each pair

- Pair (none damage, light damage):
- $\bar{x}1$ (none damage level): 18.88
- $\bar{x}2$ (light damage level): 20.32
- Point Estimate ($\bar{x}1$ - $\bar{x}2$): -1.44
- SE: 1.47
- Absolute t_score: 0.98

RESULT: FOR A P-VALUE OF PAIRS (NONE, LIGHT) = 0.164 AND ALPHA BONFERRONI = 0.00833 WE DO NOT REJECT THE NULL HYPOTHESIS!

- Pair (none damage, substantial damage):
- $\bar{x}1$ (none damage level): 18.88
- $\bar{x}2$ (substantial damage level): 27.71
- Point Estimate ($\bar{x}1$ - $\bar{x}2$): -8.83
- SE: 1.47
- Absolute t_score: 5.99

RESULT: FOR A P-VALUE OF PAIRS (NONE, SUBSTANTIAL) ~0 AND A BONFERRONI = 0.00833 WE REJECT THE NULL HYPOTHESIS!

Presentation by Ingrid Diniz, Igor Diniz and José Santos

T-TEST FOR EACH PAIR WITH BONFERRONI CORRECTION FOR ALPHA

- Pair (none damage, destructed damage):
- $\bar{x}1$ (none damage level): 18.88
- $\bar{x}2$ (destructed damage level): 25.71
- Point Estimate ($\bar{x}1 - \bar{x}2$): -6.83
- SE: 1.47
- Absolute t_score: 4.63

RESULT: FOR A P-VALUE OF PAIRS (NONE, DESTRUCTED) ~ 0 AND
A BONFERRONI = 0.00833 WE REJECT THE NULL HYPOTHESIS!

- Pair (light damage, substantial damage):
- $\bar{x}1$ (light damage level): 20.32
- $\bar{x}2$ (substantial damage level): 27.71
- Point Estimate ($\bar{x}1 - \bar{x}2$): -7.39
- SE: 1.47
- Absolute t_score: 5.01

RESULT: FOR A P-VALUE OF PAIRS (LIGHT, SUBSTANTIAL) ~ 0 AND
A BONFERRONI = 0.00833 WE REJECT THE NULL HYPOTHESIS!

T-TEST FOR EACH PAIR WITH BONFERRONI CORRECTION FOR ALPHA

- Pair (light damage, destructed damage):
- $\bar{x}1$ (light damage level): 20.32
- $\bar{x}2$ (destructed damage level): 25.79
- Point Estimate ($\bar{x}1 - \bar{x}2$): -5.38
- SE: 1.47
- Absolute t_score: 3.65

RESULT: FOR A P-VALUE OF PAIRS (LIGHT, DESTRUCTED) = 0.00014 AND A BONFERRONI = 0.00833 WE REJECT THE NULL HYPOTHESIS!

- Pair (substantial damage, destructed damage):
- $\bar{x}1$ (light damage level): 27.71
- $\bar{x}2$ (destructed damage level): 25.71
- Point Estimate ($\bar{x}1 - \bar{x}2$): 2.00
- SE: 1.47
- Absolute t_score: 1.36

RESULT: FOR A P-VALUE OF PAIRS (SUBSTANTIAL , DESTRUCTED) = 0.087 AND A BONFERRONI = 0.00833 WE DO NOT REJECT THE NULL HYPOTHESIS!

Presentation by Ingrid Diniz, Igor Diniz and José Santos

KRUSKAL-WALLIS RESULTS

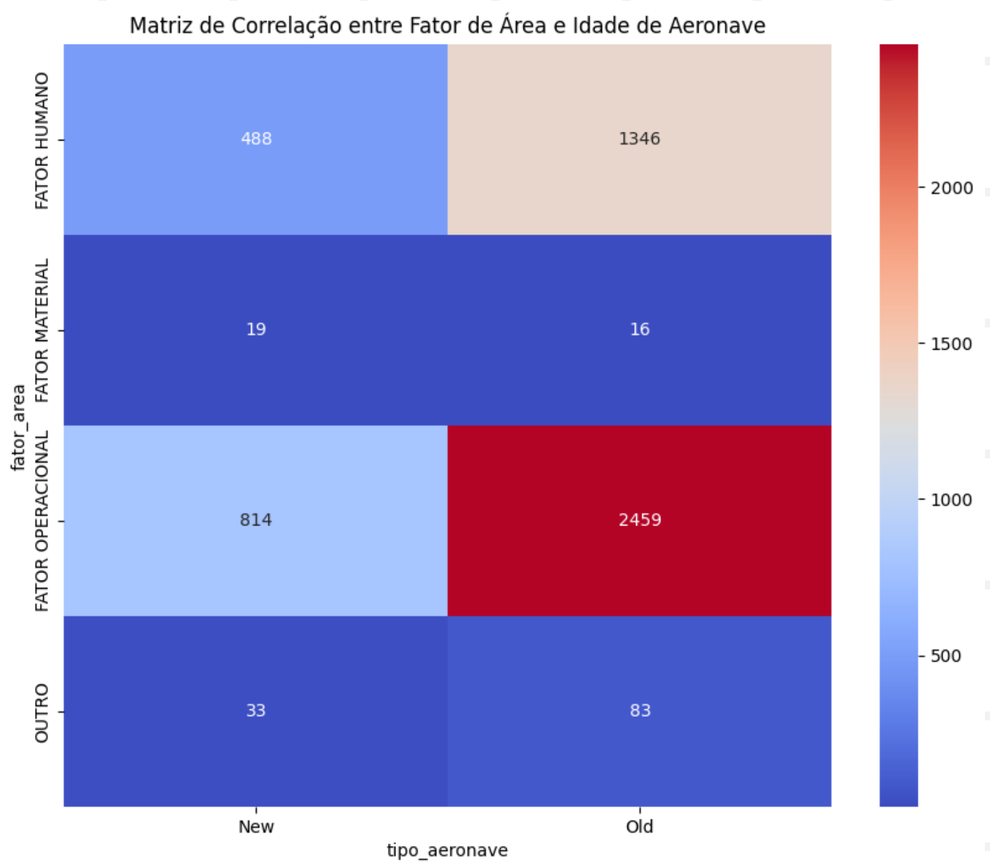
- H-score gotten from kruskal: 17.21
- p-value ~0
- α = 0.05

	DESTRUÍDA	LEVE	NENHUM	SUBSTANCIAL
DESTRUÍDA	False	True	True	False
LEVE	True	False	False	True
NENHUM	True	False	False	True
SUBSTANCIAL	False	True	True	False

RESULT: WE REJECTED THE NULL HYPOTHESIS!

Presentation by Ingrid Diniz, Igor Diniz and José Santos

IS THERE A SIGNIFICANT RELATIONSHIP BETWEEN THE OCCURRENCE FACTOR AND THE AGE OF THE AIRCRAFT?



Hypothesis Formulation

- **H₀:** There is no significant association.
- **Ha:** There is a significant association.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

IS THERE A SIGNIFICANT RELATIONSHIP BETWEEN THE OCCURRENCE FACTOR AND THE AGE OF THE AIRCRAFT?

Organization of Data in a Contingency Table

tipo_aeronave	New	Old
fator_area		
FATOR HUMANO	276	759
FATOR MATERIAL	12	13
FATOR OPERACIONAL	443	1428
OUTRO	23	46

Checking conditions

- **Independence:** Each case that contributes a count to the table is independent of all the other cases in the table and between each other.
- **Sample size/distribution:** Each particular scenario has at least 5 expected cases

Presentation by Ingrid Diniz, Igor Diniz and José Santos

IS THERE A SIGNIFICANT RELATIONSHIP BETWEEN THE OCCURRENCE FACTOR AND THE AGE OF THE AIRCRAFT?

Chi-Square Calculation

$$\chi^2 = \Sigma[(O_{ij} - E_{ij})^2 / E_{ij}]$$

Where,

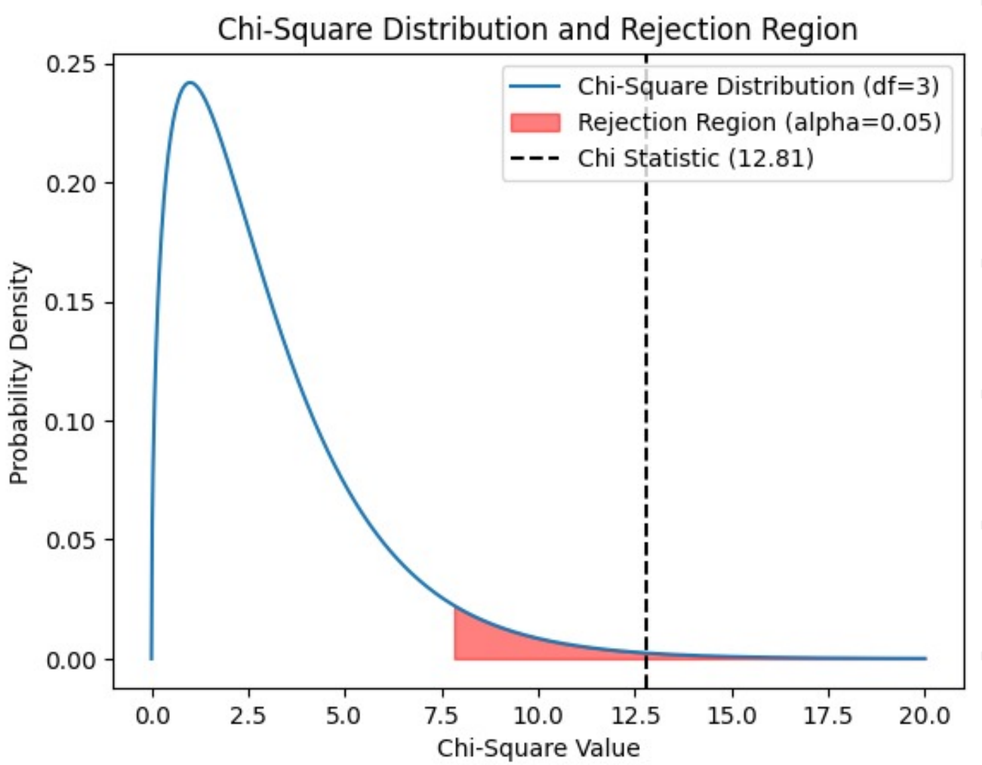
- O_{ij} is the observed number of cases in cell (i, j).
- E_{ij} is the expected number of cases in cell (i, j).

Degrees of Freedom For a Two-Way Table

$df = (R - 1) * (C - 1)$ Where,

R is the number of rows in the table and C is the number of columns.

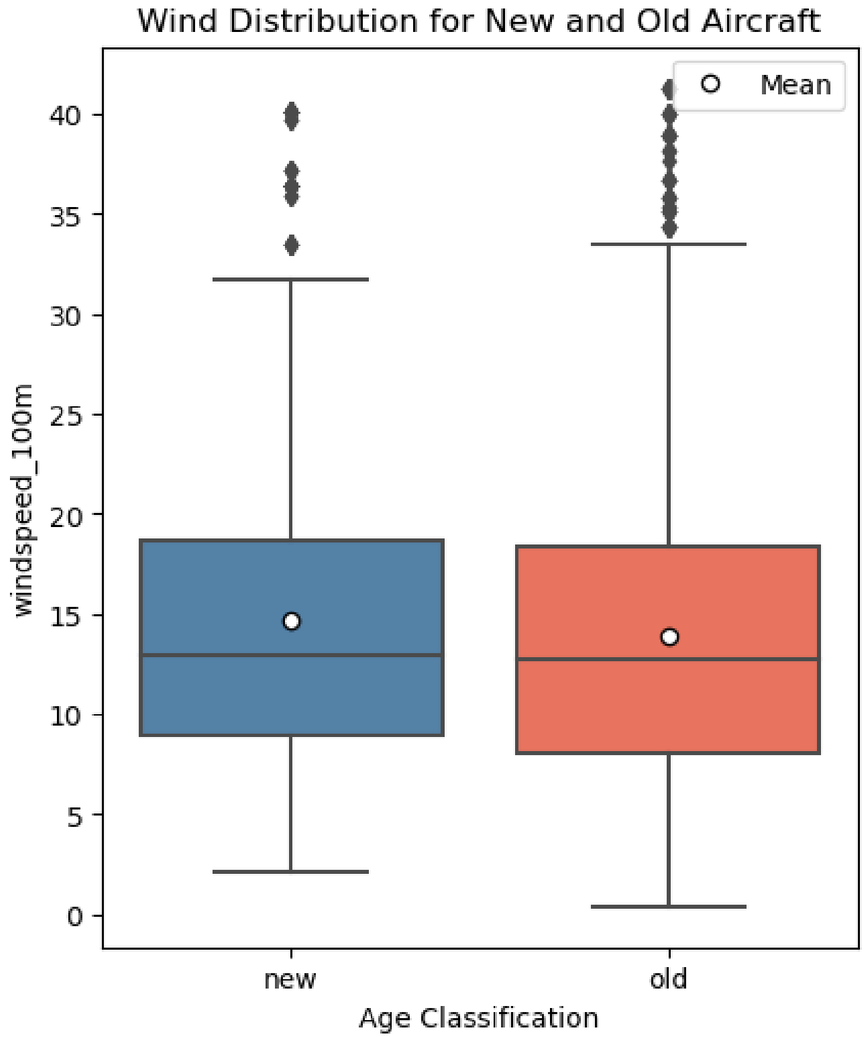
- n = 3000
- α = 0.05
- χ2: 12.81
- critical value = 7.81
- p-value: 0.005



RESULT: WE REJECTED THE NULL HYPOTHESIS!

Presentation by Ingrid Diniz, Igor Diniz and José Santos

ARE OLDER AIRCRAFT MORE SUSCEPTIBLE TO ACCIDENTS DUE TO WIND?



Hypothesis Formulation

- $H_0: \mu_{new} = \mu_{old}$
- $H_a: \mu_{new} \neq \mu_{old}$

Where μ_new represents new airplanes average and μ _old represents the old airplanes average

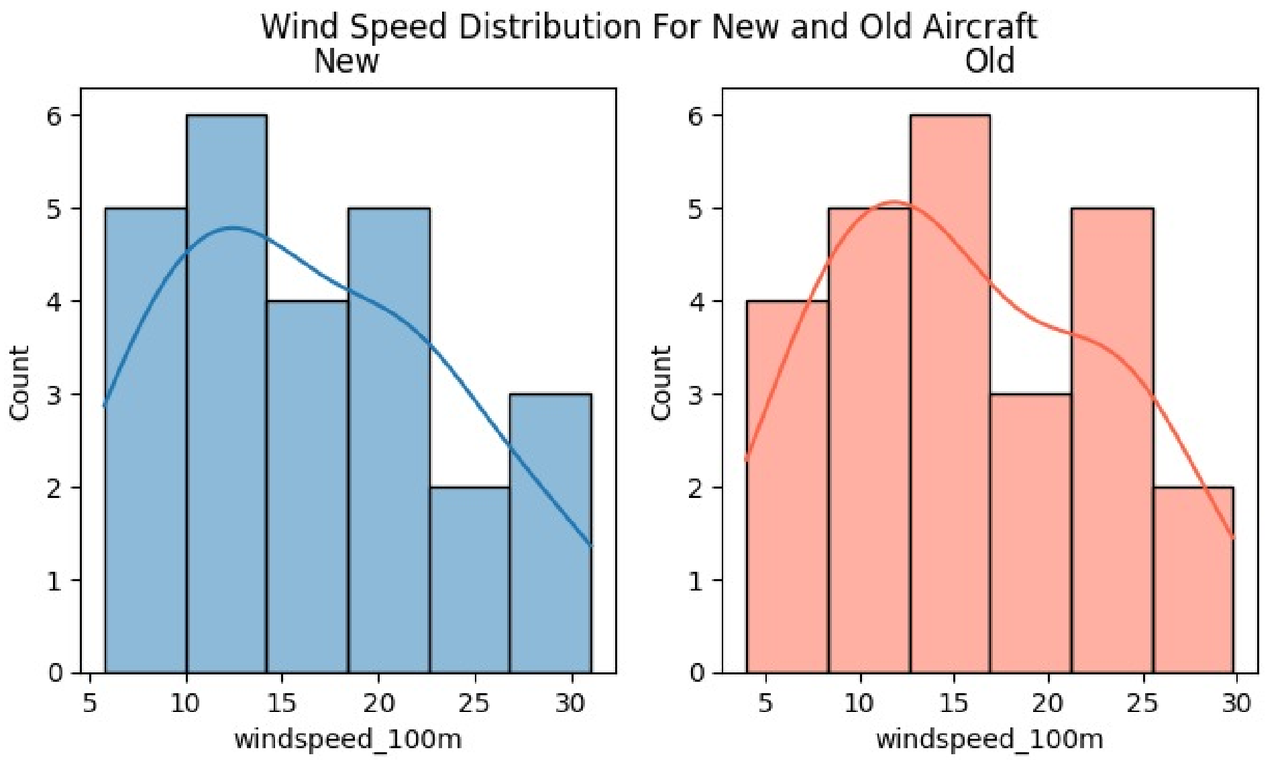
Presentation by Ingrid Diniz, Igor Diniz and José Santos

ARE OLDER AIRCRAFT MORE SUSCEPTIBLE TO ACCIDENTS DUE TO WIND?

Checking conditions

- **Independence extended:** The data are independent within and between the two groups since we are taking simple random samples.

- **Normality:** We check the outliers rules of thumb for each group separately and there are no clear outliers in the data.



Presentation by Ingrid Diniz, Igor Diniz and José Santos

ARE OLDER AIRCRAFT MORE SUSCEPTIBLE TO ACCIDENTS DUE TO WIND?

Standard Error Calculate

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$point\ estimate = \bar{x}_{new} - \bar{x}_{old} = 0.78$$

Confidence Interval

$$CI = point\ estimate\ \pm\ t^*_{df}\ *\ SE$$

- n1 = n2 = 25
- α = 0.05
- std_new = 7.18
- std_old = 7.06
- SE = 2.015
- t-test = 2.06
- (-3.37, 4.94)

RESULT: CONFIDENCE INTERVAL = (-3.37, 4.94)

Presentation by Ingrid Diniz, Igor Diniz and José Santos

RESULTS

In the “difference between the proportion of accidents in older aircraft than newer aircraft” analysis, since we have rejected the null hypothesis, we can conclude that there is a difference in the proportion of occurrences classified as **accidents** between older aircraft and newer ones.

Furthermore, based on the confidence interval, we can also confirm with a 95% confidence level that older aircraft tend to have between 6.1 to 23.23 percentage points greater number of accidents than newer ones.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

RESULTS

In the “difference among the averages of airplanes in each damage level” analysis, after applying the BONFERRONI correction and doing a t-test for each pair of damage levels using the corrected α and df (degree of freedom), we can conclude that there is a significant difference in the average of age of airplanes between the following pairs of damage level:

- None vs Substantial

We can even conclude that the direction of the difference between them is that the average of airplanes with **substantial damage** is greater than the ones with **no damage** according to the calculated point estimate.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

RESULTS

- None vs Destructed

We can even conclude that the direction of the difference between them is that the average of airplanes with **destructed damage** is greater than the ones with **no damage** according to the calculated point estimate.

- Light vs Substantial

We can even conclude that the direction of the difference between them is that the average of airplanes with **substantial damage** is greater than the ones with **light damage** according to the calculated point estimate.

- Light vs Destructed

We can even conclude that the direction of the difference between them is that the average of airplanes with destructed damage is greater than the ones with light damage according to the calculated point estimate.

RESULTS

- Regarding the other pairs **(Substantial, Destructed)** and **(Light, None)**, as we have not rejected the null hypothesis, we can conclude nothing, neither if there is nor there is no difference between their averages of airplane ages.

  We also have confirmed the same results using the Kruskal-Wallis test followed by pairwise Dunn's test, which is more appropriate when the normality condition is not respected.

RESULTS

- In the “ significant relationship between the occurrence factor and the age of the aircraft” analysis, since we have rejected the null hypothesis, we can conclude that there is a significant relationship between the age of the aircraft and the factor of occurrences. And, analyzing the heatmap, we can see in which pairs there’s a stronger relationship.
- In the “aircraft more susceptible to accidents due to wind” analysis, Since the confidence interval **contains the 0 value,** we CANNOT reject with a 95.0% confidence level that the means of the wind speed, when the occurrences had happened, for older and newer aircraft are different.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

CONCLUSION

In answer to our research question, "Is the age of an aircraft related to occurrences?", our analysis provides compelling evidence that aircraft age does indeed influence the of occurrences.

Presentation by Ingrid Diniz, Igor Diniz and José Santos

THANK YOU !

Presentation by Ingrid Diniz, Igor Diniz and José Santos