# TITANIC DATA ANALYSIS

SANIKA SAWANT

DURGESH SHINDE

SAARTHAK MAHAJAN

KARTIK WATEGAONKAR

YASH YEOLA

**Guided By- Sanket B**
**Github ID : bsanketm**

# Objectives

- To perform the Exploratory Data Analysis of the given Titanic Dataset.
- Segregation of passengers on basis of gender, Cabin class, age.
- To analyse the ticketing pattern in the given data.
- To analyse the patterns of rescue of survivors.

# ABOUT THE DATASET

- **What is EDA?**
  Exploratory Data Analysis (EDA) is a method used to analyze and summarize datasets. Majority of the EDA techniques involve the use of graphs.
- **Titanic Dataset –**
  It is one of the most popular datasets used for understanding statistical analysis basics. It contains information of all the passengers aboard the RMS Titanic, which unfortunately was shipwrecked. This dataset can be used to predict whether a given passenger survived or not.

# INFORMATION ABOUT VARIABLES

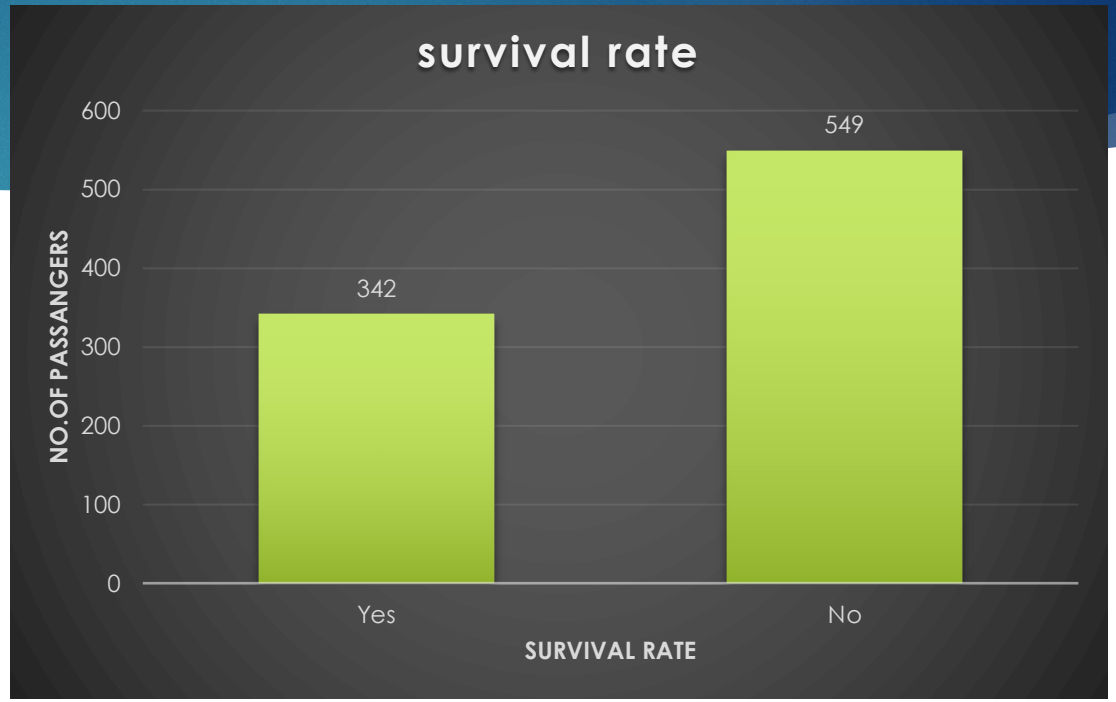
1. PassengerId: Unique Id of a passenger
2. Survived: If the passenger survived(0-No, 1-Yes)
3. Pclass: Passenger Class (1 = 1st, 2 = 2nd, 3 = 3rd)
4. Name: Name of the passenger
5. Sex: Male/Female
6. Age: Passenger age in years
7. SibSp: No of siblings/spouses aboard
8. Parch: No of parents/children aboard
9. Ticket: Ticket Number
10. Fare: Passenger Fare
11. Cabin: Cabin number
12. Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

# Types of features:

- CATEGORICAL/NOMINAL: VARIABLES THAT CAN BE DIVIDED INTO MULTIPLE CATEGORIES BUT HAVING NO ORDER OR PRIORITY.
  EG. EMBARKED (C = CHERBOURG; Q = QUEENSTOWN; S = SOUTHAMPTON)
- BINARY: A SUBTYPE OF CATEGORICAL FEATURES, WHERE THE VARIABLE HAS ONLY TWO CATEGORIES.
  EG: SEX (MALE/FEMALE)
- ORDINAL: THEY ARE SIMILAR TO CATEGORICAL FEATURES BUT THEY HAVE AN ORDER. (I.E. CAN BE SORTED).
  EG. PCLASS (1, 2, 3)
- CONTINUOUS: THEY CAN TAKE UP ANY VALUE BETWEEN THE MINIMUM AND MAXIMUM VALUES IN A COLUMN.
  EG. AGE, FARE
- COUNT: THEY REPRESENT THE COUNT OF A VARIABLE.
  EG. SIBSP, PARCH
- USELESS: THEY DON'T CONTRIBUTE TO THE FINAL OUTCOME OF AN ML MODEL.
  HERE, PASSENGERID, NAME, CABIN AND TICKET MIGHT FALL INTO THIS CATEGORY.

# Passengers survived

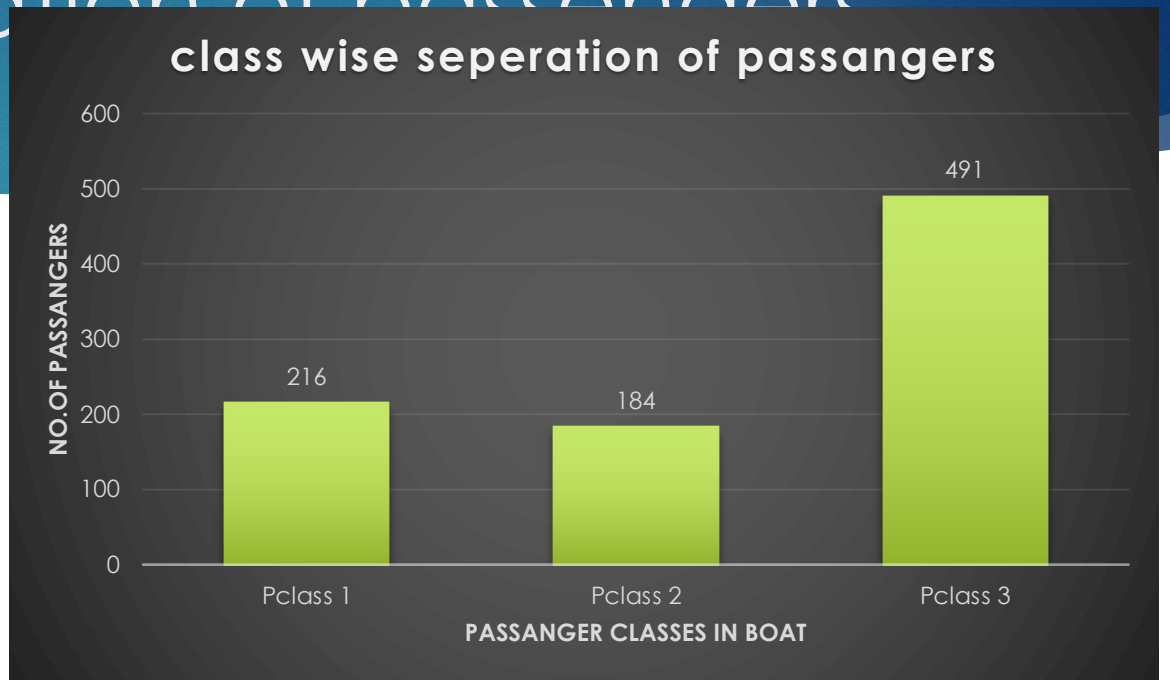
| Yes | No |
|---|---|
| 342 | 549 |




This clustered column chart shows the data about the number of passengers who survived on the board.
In the given data of 891 passengers 342 passengers survived.
The column chart helps us to by sorting the data on the basis of survival which further helps us to analyze the data.

# Class wise separation of passengers

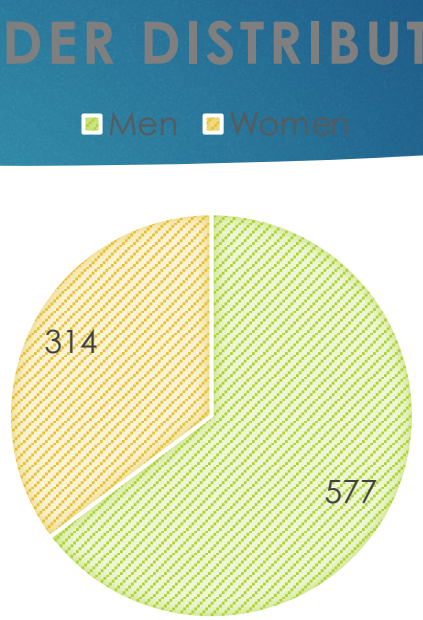
| | | |
|---|---|---|
| Pclass 1 | Pclass 2 | Pclass 3 |
| 216 | 184 | 491 |




This clustered column chart separates the data of passengers of class 1, class 2 and class 3.
From this chart it can be seen that there are more passengers belonging to class 3.

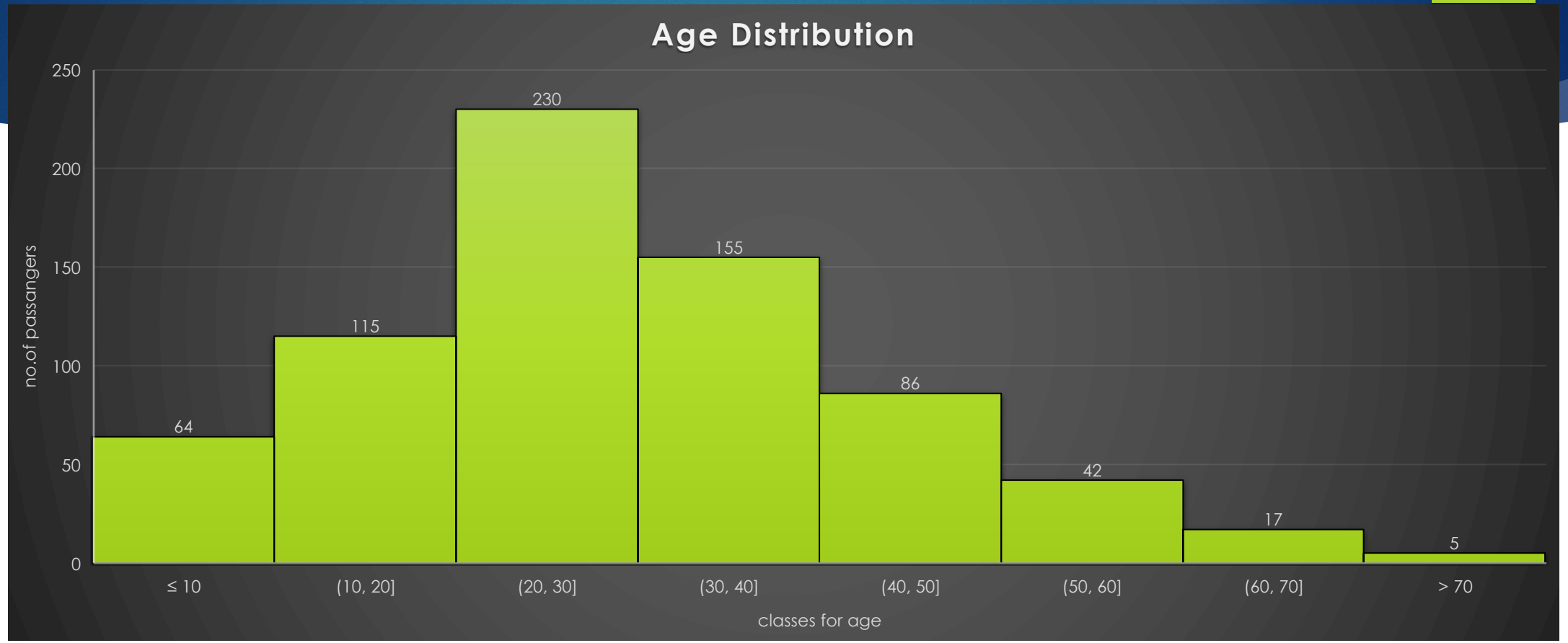
# Gender distribution

| Men | Women |
| --- | --- |
| 577 | 314 |




This pie chart tell us about the distribution of passengers on the basis of gender.
There are 577 males and 314 females.

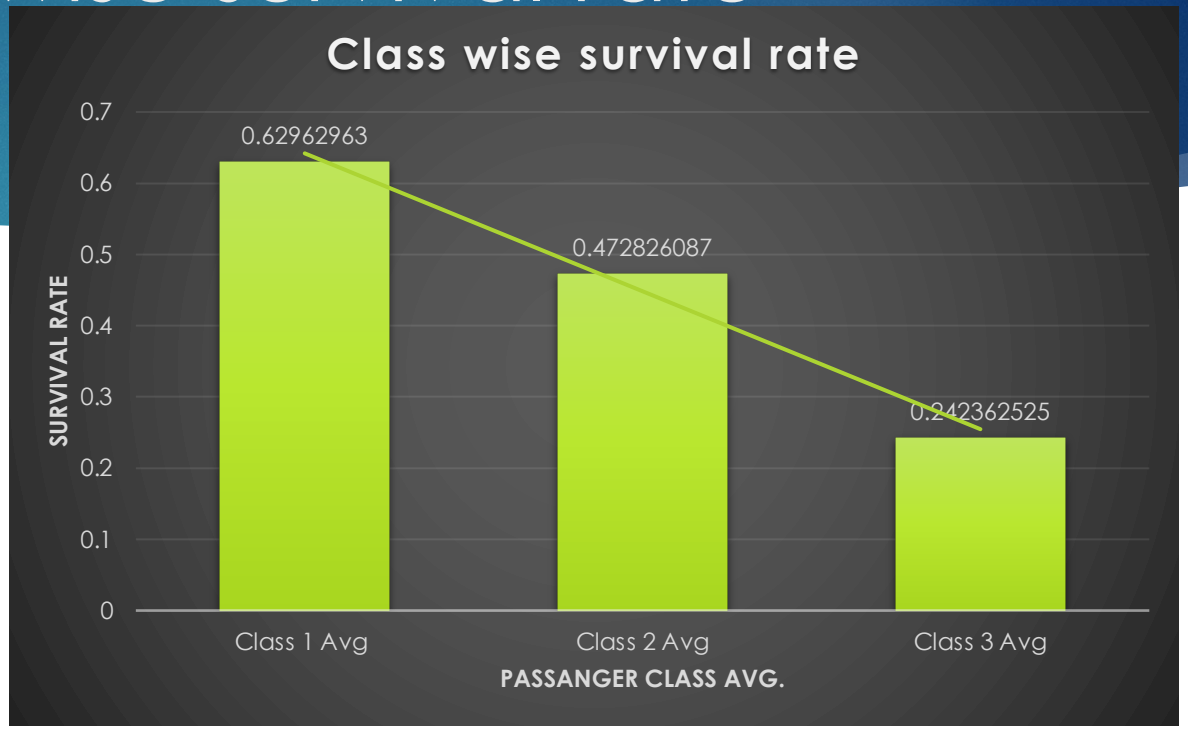
# Age Distribution




This histogram shows frequencies of number of passengers in different age group. From the above histogram it can be seen that maximum passengers belong to the age group of 20-30 followed by age group of 30-40.

# Class wise survival rate

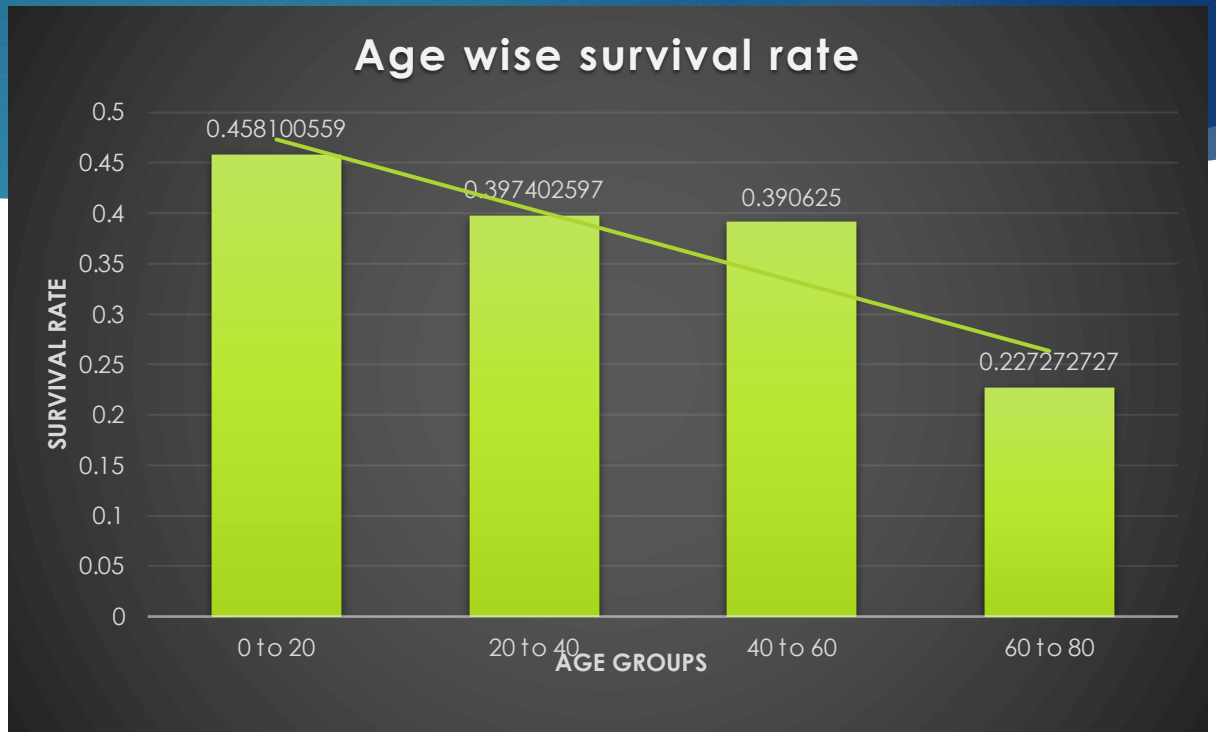
| Class 1 Avg | Class 2 Avg | Class 3 Avg |
|---|---|---|
| 0.62962963 | 0.472826087 | 0.242362525 |




From the above chart it can be clearly analyzed that more preference was given to the passengers of class 1 while rescuing as compared to passengers of class 2 to class 3.
It can be inferred that survival rate of class 1 passengers is highest while that of class 3 passengers is the least.

# Age wise survival rate

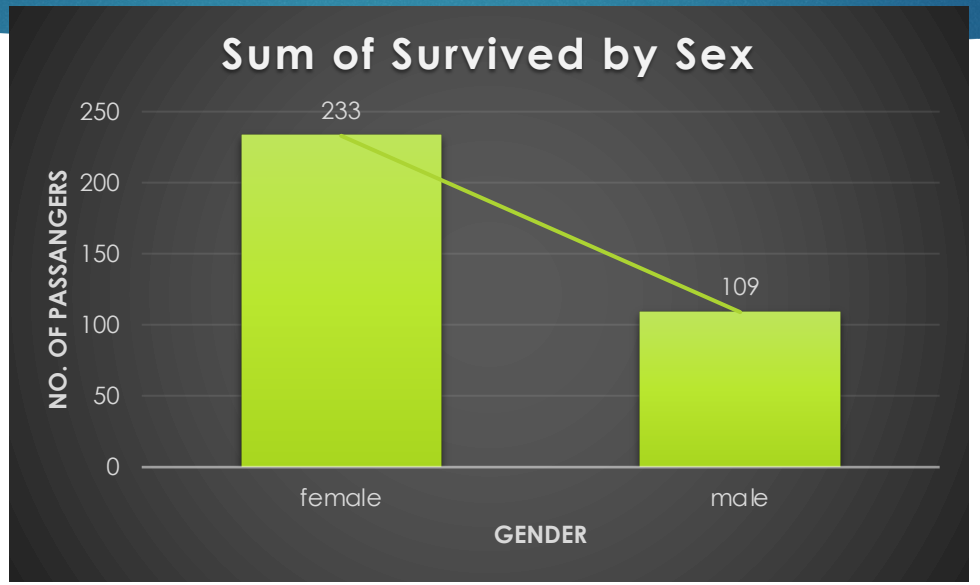
| 0 to 20 | 20 to 40 | 40 to 60 | 60 to 80 |
|---|---|---|---|
| 0.458101 | 0.397403 | 0.390625 | 0.227273 |




From the above graphical representation it can be concluded that priorities were given to younger passengers when it came to rescuing process of passengers.

# Sex wise survival

| Sex | Sum of Survived |
|---|---|
| female | 233 |
| male | 109 |




Above chart tells us that, out of 891 passengers only 342 survived.
Among those 342 , 233 passengers are female.
It can be inferred that while rescue operations more females were preferred over males.

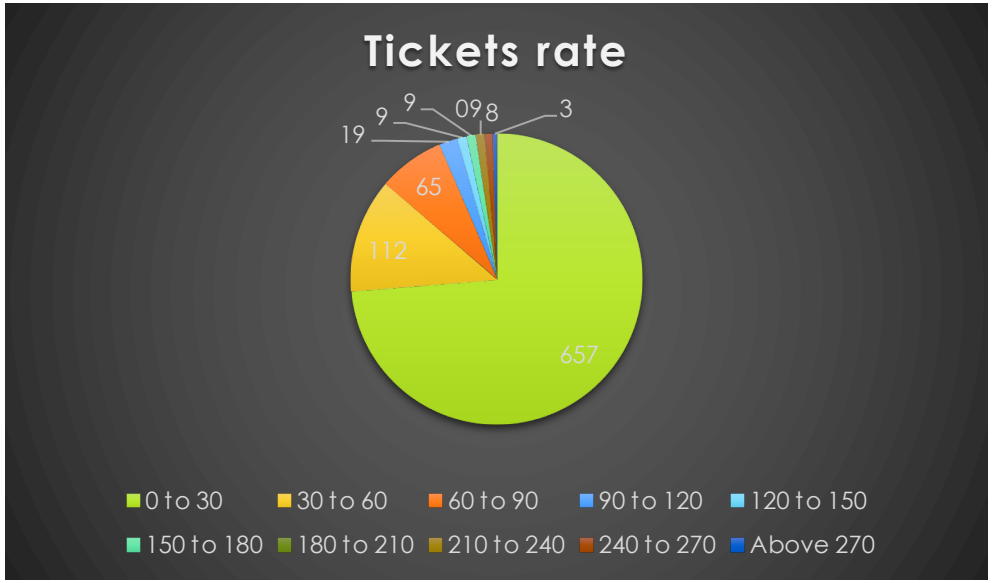
# Descriptive statistics on ticket fares

| | Tickets price summary statistics | | |
|---|---|---|---|
| | | | |
| | Mean | 32.2322464 | |
| | Standard Error | 1.666427709 | |
| | Median | 14.4542 | |
| | Mode | 8.05 | |
| | Standard Deviation | 49.71431749 | |
| | Sample Variance | 2471.513364 | |
| | Kurtosis | 33.36721986 | |
| | Skewness | 4.785121639 | |
| | Range | 512.3292 | |
| | Minimum | 0 | |
| | Maximum | 512.3292 | |
| | Sum | 28686.6993 | |
| | Count | 890 | |
| | | | |

Average of ticket fares is 32.2322464, which means that there are more passengers sailing in class 2.
Total amount paid by all the passengers on board is 28686.6993.
The minimum ticket fare is 0 while maximum ticket fare is 512.3292.

# Tickets Rate

| 0 to 30 | 30 to 60 | 60 to 90 | 90 to 120 | 120 to 150 | 150 to 180 | 180 to 210 | 210 to 240 | 240 to 270 | Above 270 |
|---|---|---|---|---|---|---|---|---|---|
| 657 | 112 | 65 | 19 | 9 | 9 | 0 | 9 | 8 | 3 |




The inference that can be drawn from the above pie chart is that there are maximum passengers buying ticket in the range of 0-30 $.

# Correlation

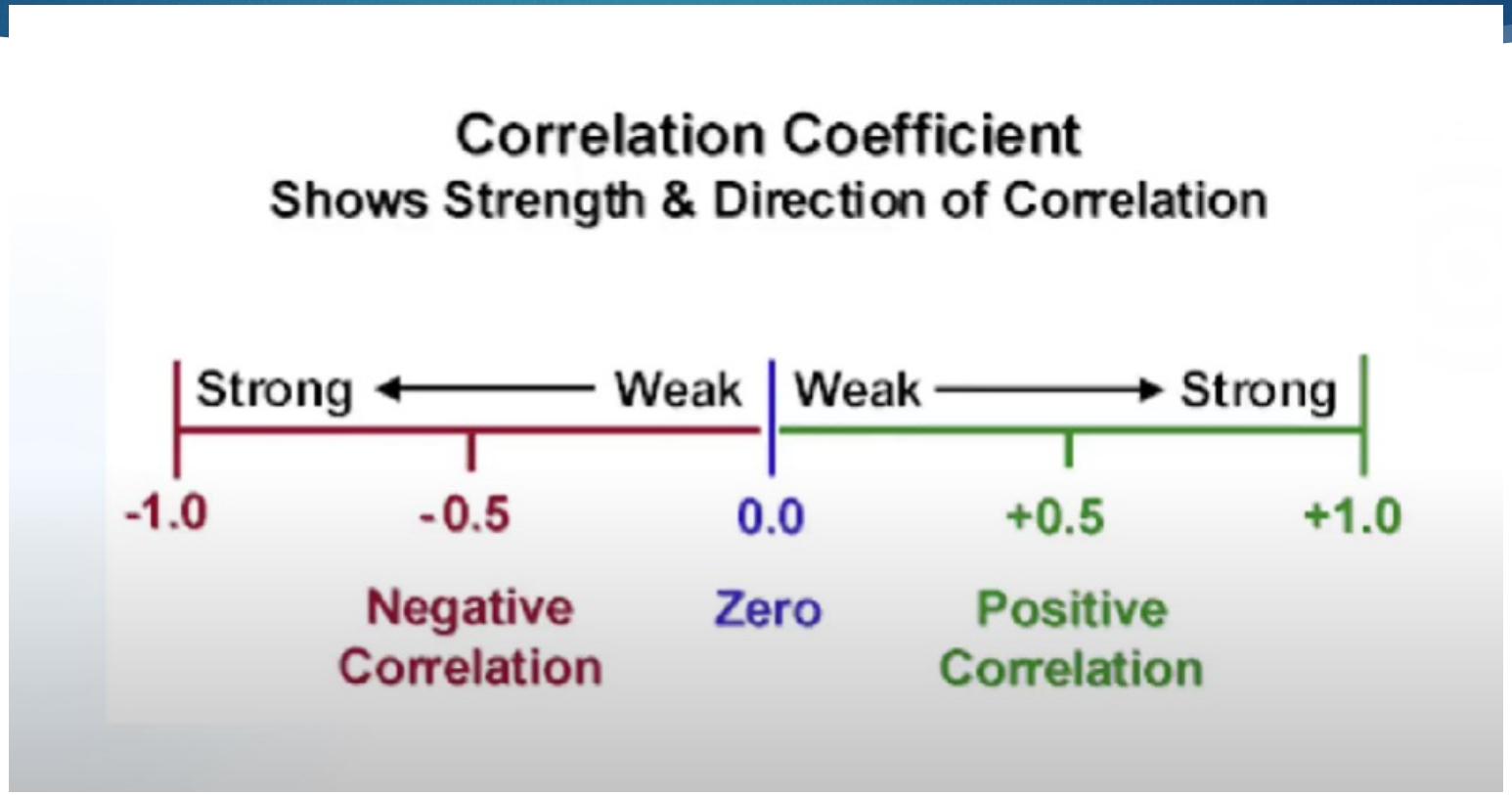
- In statistics, dependence is any statistical relationship between two random variables or two sets of data.
- Correlation refers to any of a broad class of statistical relationships involving dependence. Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price.

# Positive & Negative Correlation

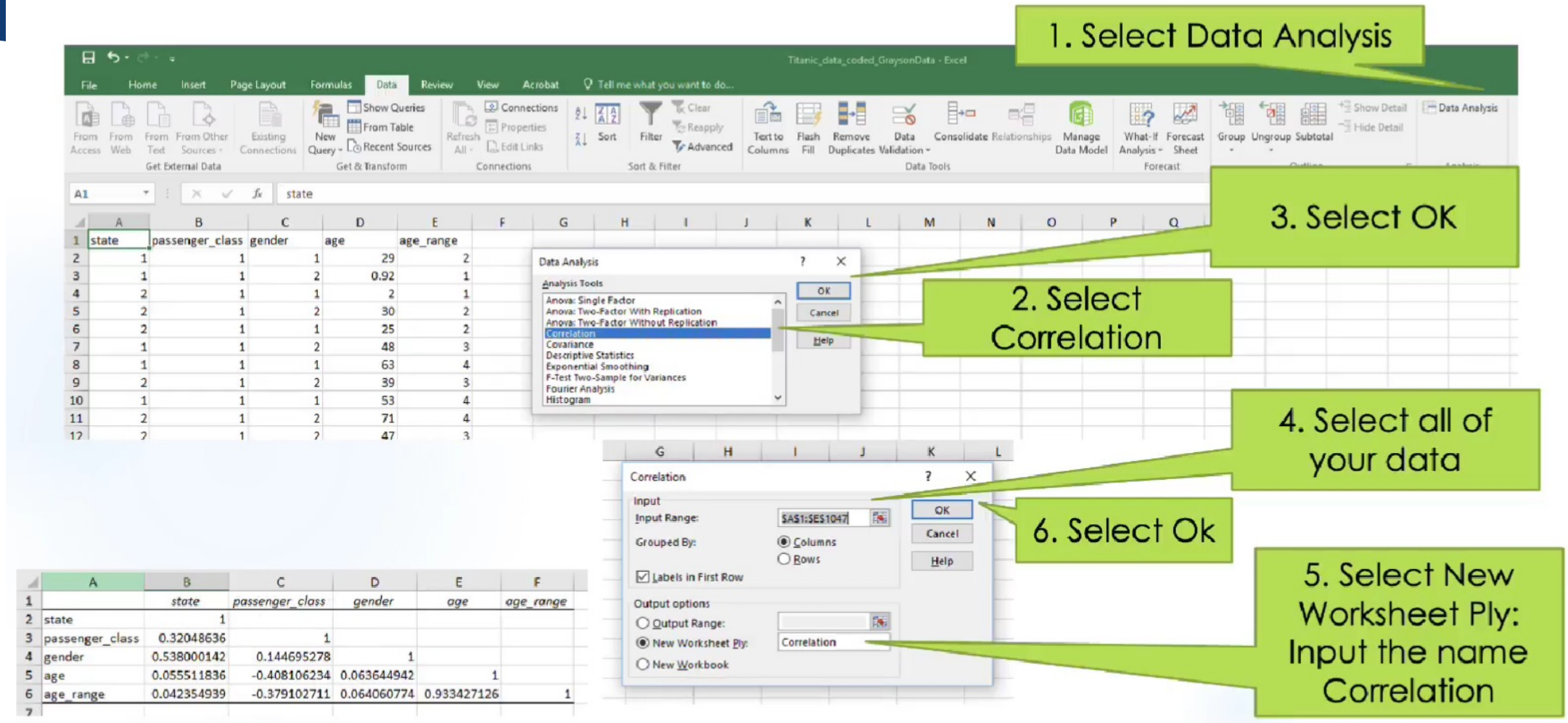
- The two directions of a correlation are positive and negative.
- In a positive correlation, both variables move in the same direction. In other words, as one yarlable increases, so does the other. for example, there is a positive correlafion between smoking and alcohol use. As alcohol use increases, so does smoking.
- When two variables have a negative correlation, they have an inverse relationship.This means that as one variable increases, the other decreases, and vice versa.Negative correlations are indicated by a minus (- sign in front of the correlation value.

- Students who spent the higher amount of time playing video games each week had the lowest GPAs. As the hours spent playing video games decreased. the GPAs increased.
- The weight of a car and miles per gallon: cars that are heavier tend fo get less miles per gallon of gas.
- School achievement and days absent from school: people who miss more days of school fend fo have lower GPAs.Vaccinations and illness: The more that people are vaccinated for a specific illness. the less that illness occurs.

# CORRELATION STRENGTH



# CORRELATION-PEARSON'S R TEST

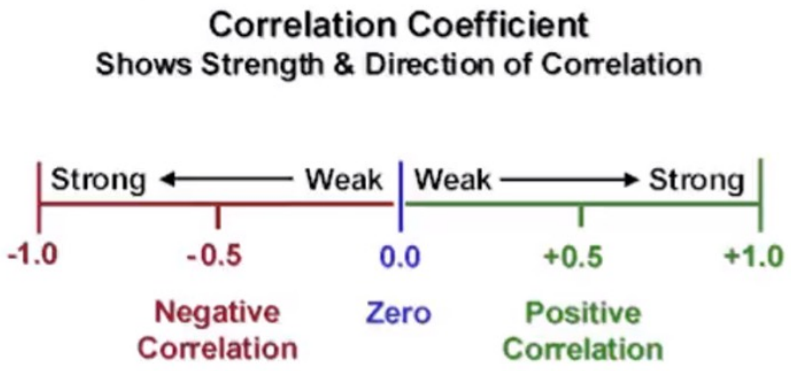



| | state | passenger_class | gender | age | age_range |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 29 | 2 |
| 2 | 1 | 1 | 2 | 0.92 | 1 |
| 3 | 2 | 1 | 1 | 2 | 1 |
| 4 | 2 | 1 | 2 | 30 | 2 |
| 5 | 2 | 1 | 1 | 25 | 2 |
| 6 | 1 | 1 | 2 | 48 | 3 |
| 7 | 1 | 1 | 1 | 63 | 4 |
| 8 | 2 | 1 | 2 | 39 | 3 |
| 9 | 1 | 1 | 1 | 53 | 4 |
| 10 | 2 | 1 | 2 | 71 | 4 |
| 11 | 2 | 1 | 1 | 47 | 3 |

| | | state | passenger_class | gender | age | age_range |
|---|---|---|---|---|---|---|
| 1 | | 1 | | | | |
| 2 | state | 0.32048636 | 1 | | | |
| 3 | passenger_class | 0.538000142 | 0.144695278 | 1 | | |
| 4 | gender | 0.055511836 | -0.408106234 | 0.063644942 | 1 | |
| 5 | age | 0.042354939 | -0.379102711 | 0.064060774 | 0.933427126 | 1 |
| 6 | age_range | | | | | |

# Testing Correlation Strength

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | *state* | *passenger_class* | *gender* | *age* | *age_range* |
| 2 | state | 1 | | | | |
| 3 | passenger_class | 0.32048636 | 1 | | | |
| 4 | gender | 0.538000142 | 0.144695278 | 1 | | |
| 5 | age | 0.055511836 | -0.408106234 | 0.063644942 | 1 | |
| 6 | age_range | 0.042354939 | -0.379102711 | 0.064060774 | 0.933427126 | 1 |
| 7 | | | | | | |



# Regression

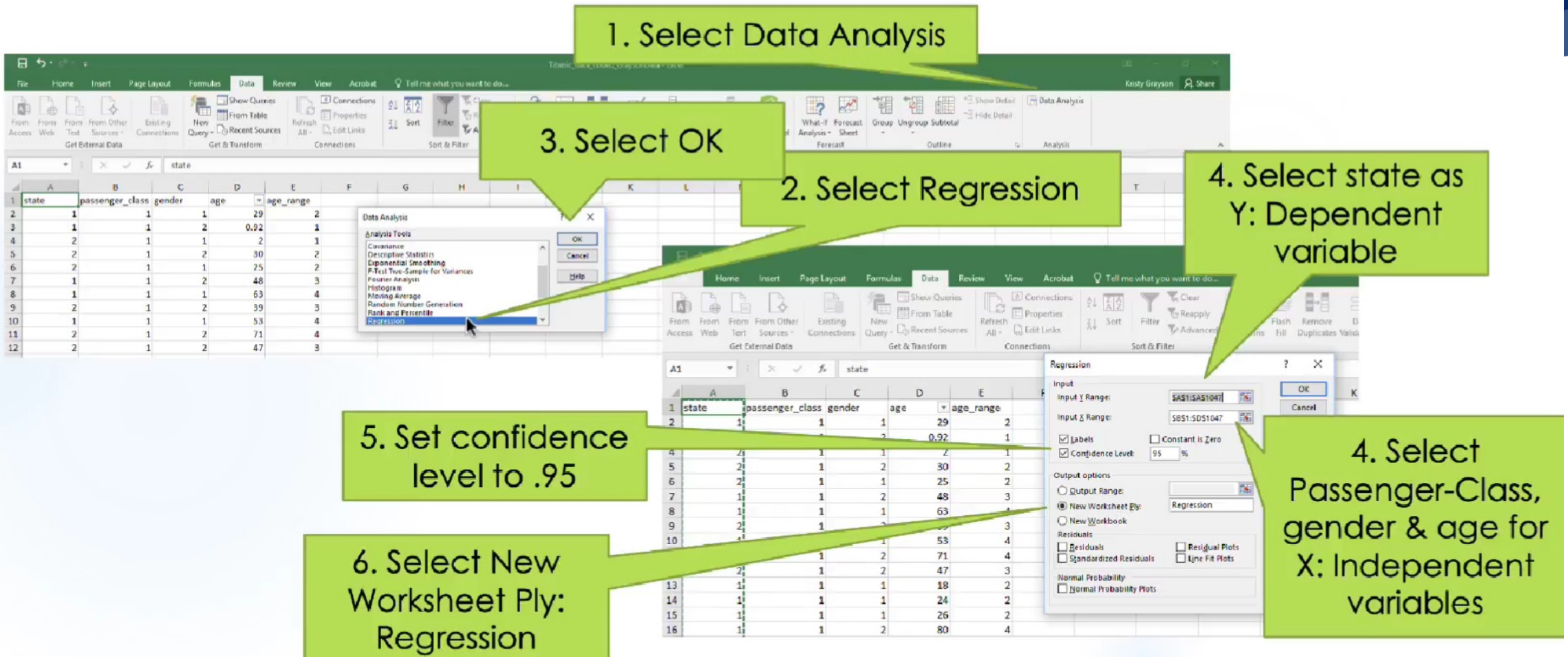
- In statistical analysis, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features').
- We attempt to predict the factors that are influencing an outcome.
- Ex: What factors are influencing this months sales?

Price, promotions, placement of ads, etc.

Dependent variable = Sales

Independent variables = price, promos, placement

# Analyzing our data using regression



# Regression Model and ANOVA

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.607110965 |
| R Square | 0.368583724 |
| Adjusted R Square | 0.366765827 |
| Standard Error | 0.391306846 |
| Observations | 1046 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 3 | 93.13716042 | 31.04572014 | 202.7527928 | 1.4444E-103 |
| Residual | 1042 | 159.5521321 | 0.153121048 | | |
| Total | 1045 | 252.6892925 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 0.232230965 | 0.062184835 | 3.734527321 | 0.000198256 | 0.110209193 | 0.354252737 | 0.110209193 | 0.354252737 |
| passenger_class | 0.182787643 | 0.016040729 | 11.39522028 | 1.96741E-28 | 0.151311831 | 0.214263456 | 0.151311831 | 0.214263456 |
| gender | 0.491487446 | 0.025550102 | 19.23622219 | 8.36717E-71 | 0.441351931 | 0.541622962 | 0.441351931 | 0.541622962 |
| age | 0.005200181 | 0.000928526 | 5.600469462 | 2.73403E-08 | 0.003378187 | 0.007022174 | 0.003378187 | 0.007022174 |