2025

4Geeks Academy: data science cohort 12

DAY 11: DESCRIPTIVE STATISTICS

TODO

DESCRIPTIVE STATISTICS	Common summary statistics & interpretation
PROBABILITY PROJECT	Submit Probability Exercises in Python Project (Probability module), if you haven't done so already
DESCRIPTIVE STATS PROJECT	Work on Descriptive Statistics Exercises in Python Project (Descriptive Stats module), plan to finish before class Monday

Introduction & basics
Weeks 1 - 5

Background needed to think like a data scientist

1. ~~Pandas~~
2. ~~Data visualization~~
3. ~~Intro to SQL~~
4. ~~Web Scraping~~
5. ~~API requests~~
6. ~~Calculus and L.Algebra~~
7. ~~Probability~~
8. Descriptive Statistics
9. Random Variables
10. Hypothesis Testing
11. Algorithm Optimization

Data science tools & techniques
Weeks 6 - 10

Data science concepts & techniques

1. Exploratory analysis
2. Your first ML algorithm
3. Linear regression
4. Decision tree algo
5. Random forest
6. Boosting algorithms
7. Naive bayes algorithm
8. K-nearest neighbors
9. Unsupervised learning
10. Time series forecasting
11. Intro to deep learning
12. Deep learning
13. Intro to NLP
14. Recommendation systems
15. ML web app using Flask
16. ML web app using Streamlit
17. Cloud computing for ML

Final project
Week 11 - 16

End-to-end data science application

1. Small dev teams
2. You pick the topic
3. Build and deploy an application
4. Pitch it on GeekTalk day

Past project topics:

1. Aviation incident prediction
2. Workout assistant
3. Cancer diagnosis
4. Natural disaster forecasting
5. Fantasy sports assistant

TOPICS

01 CENTRAL TENDENCY

02 SHAPE & DISPERSION

CENTRAL TENDENCY

WHAT

Describes a dataset's values with one number

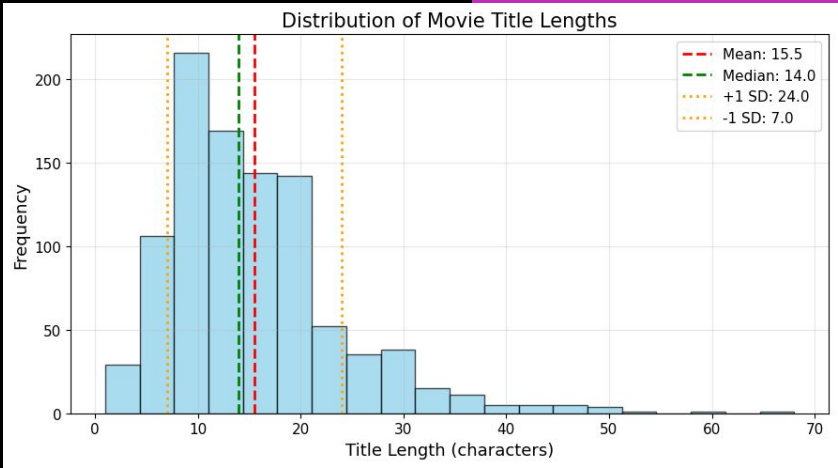
MEAN

- Numpy & Pandas `.mean()` method
- Average value of the variable

$$mean = \frac{sum of values}{count}$$

MEDIAN

- Numpy & Pandas `.median()` method
- Middle value when data is sorted
- Less sensitive than mean to skew & outliers



SHAPE & DISPERSION

WHAT

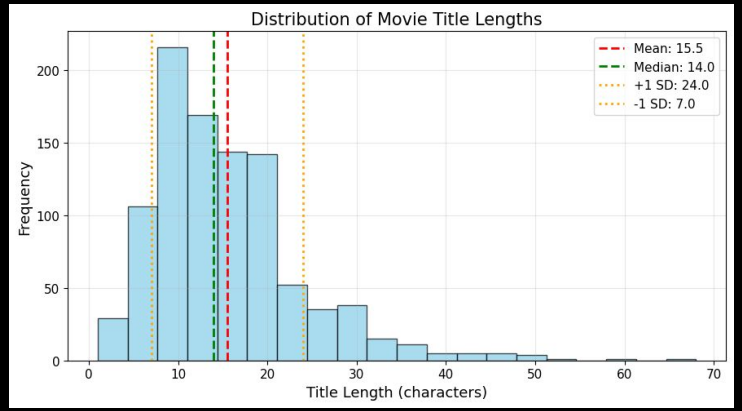
Describes a dataset's shape and spread

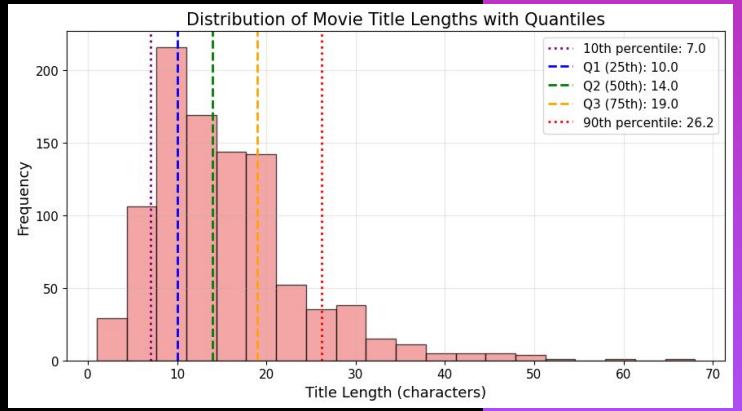
STANDARD DEVIATION

- Numpy & Pandas `.std()` method
- Average spread of the data around the mean

QUANTILES

- Numpy & Pandas `.percentile()` method
- Percentiles: how much data is below a given percentile
- Quantile: divides data into four equal parts (boxplot)





BAYESIAN PROBABILITY

WHAT

Treats probability is an expectation given a state of knowledge

WHY

Allows for the inclusion of prior knowledge and updating of probability based on new knowledge (**Bayesian inference**)

HOW

Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

EXAMPLE

A model is trained to detect spam on a dataset of 100,000 emails, 1000 of which are spam. Of flagged emails, 95% are actually spam. It also incorrectly identifies 1% of non-spam emails as spam. If the model flaggs a new email as spam, what is the probability that it really is spam?

There is only a 2.4% chance the new email is spam!

$$P(flagged|spam) = \frac{P(spam|flagged) \times P(spam)}{P(flagged)}$$

$$P(flagged|spam) = \frac{0.95 \times 0.050}{0.019} = 0.024$$