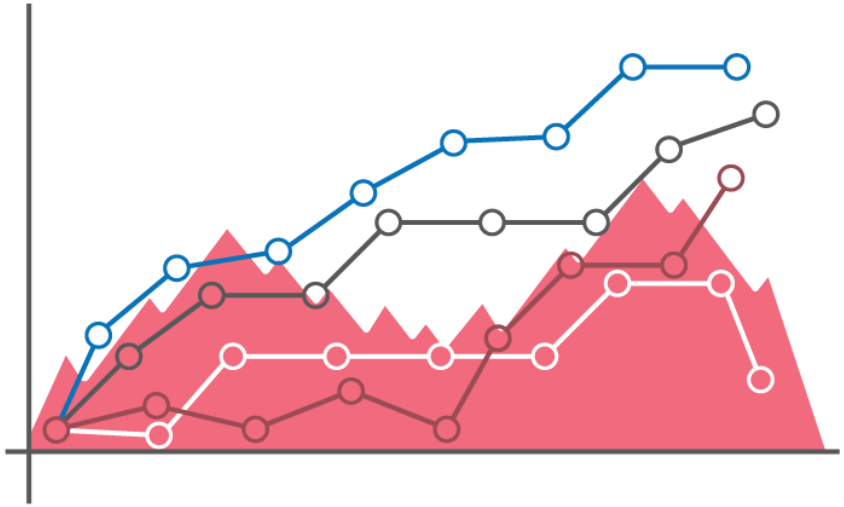
edureka!



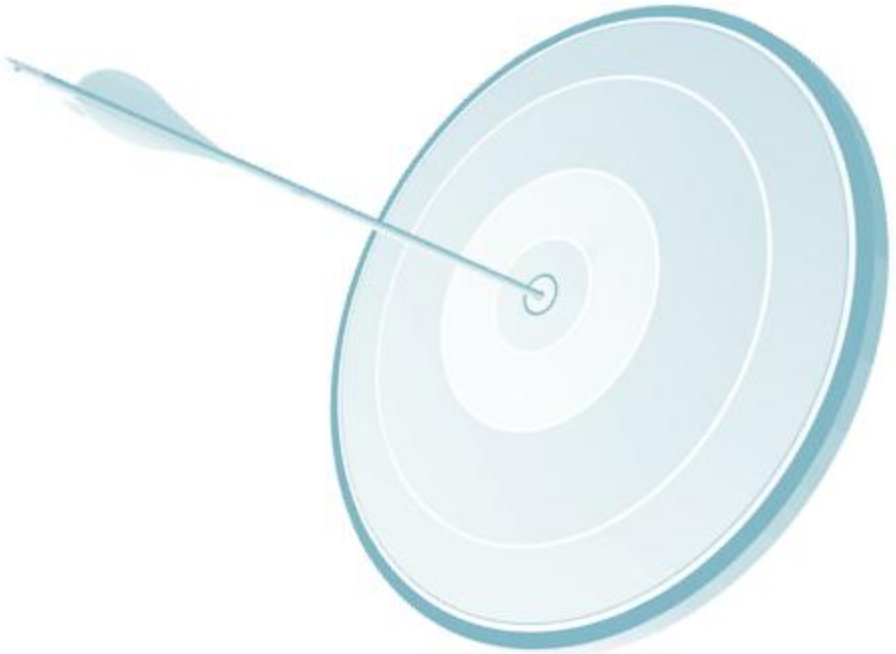
MODULE-3
DISTRIBUTIONS AND REGRESSION MODELING

Objectives

edureka!

At the end of this module, you will be able to:

- → Understand what is Normal distribution
- → Interpreting z-scores and calculating percentiles
- → Binomial Distribution
  - » Definition, properties, conditions
  - » Calculating probabilities
  - » Mean and standard deviation
- → Understand the Milgram Experiment



Course Topics

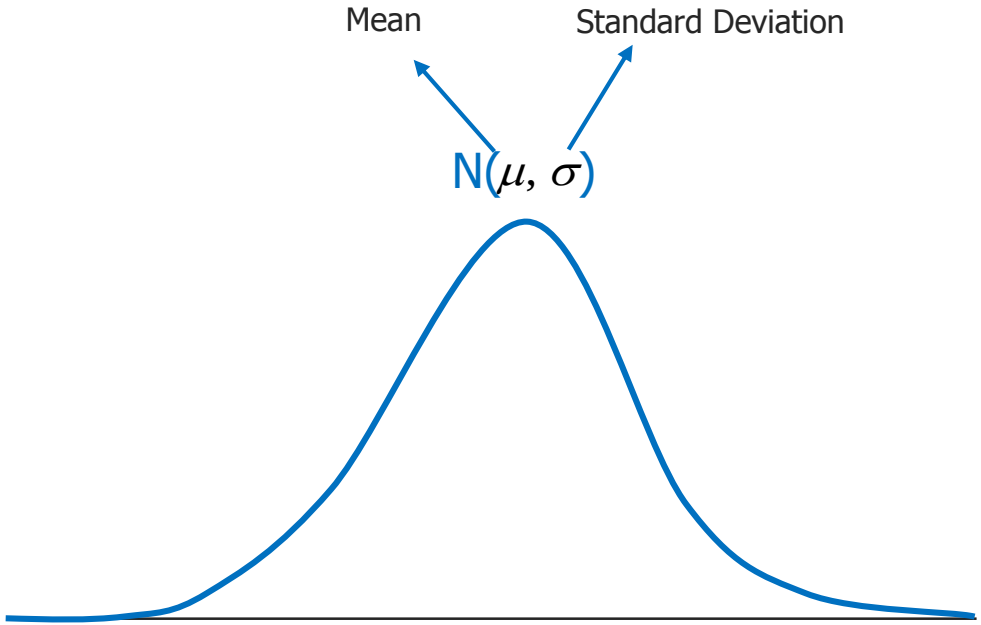
edureka!

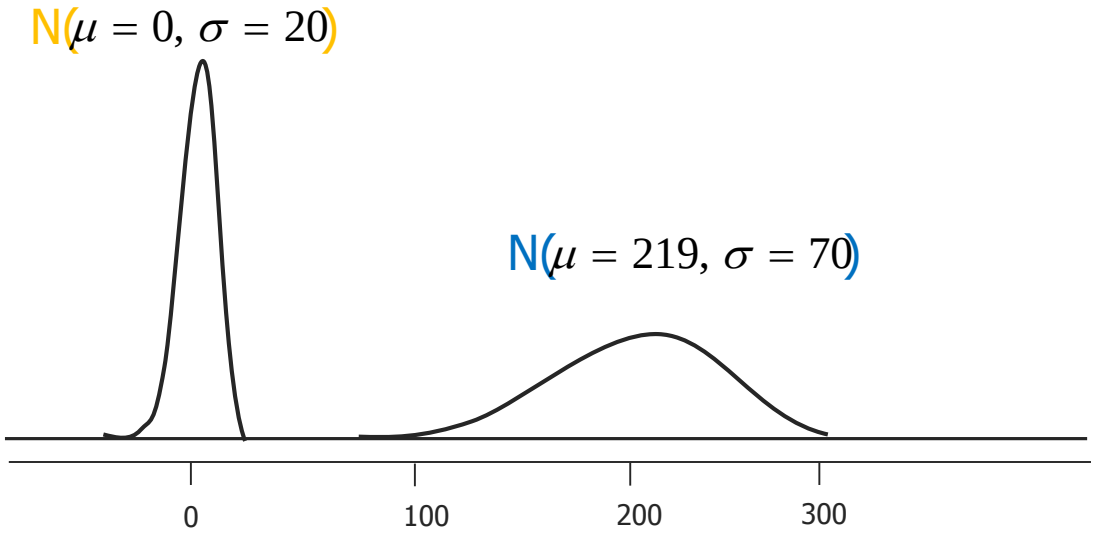
- Module 1
  - Statistics and Basic Probability
- Module 2
  - Conditional Probability and Bayesian Inference
- Module 3
  - **Probability Distributions and Regression Modelling**

Normal Distribution

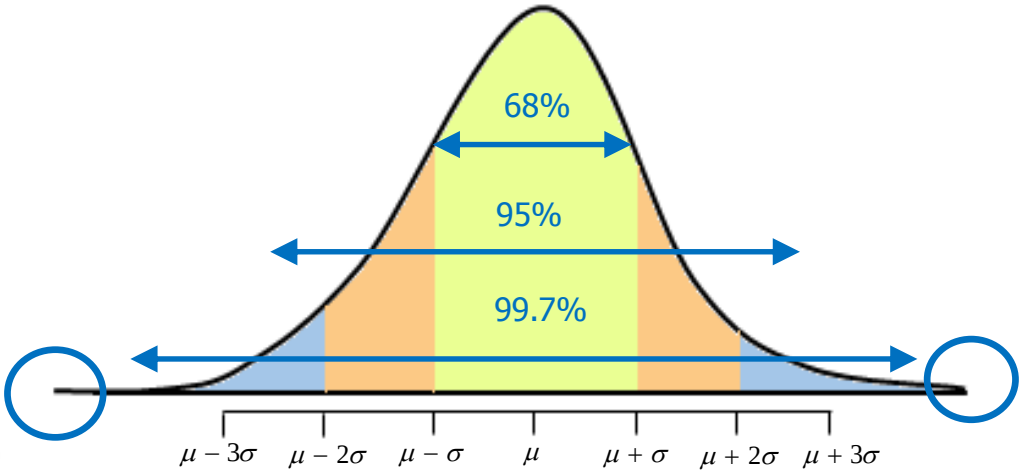
→ Unimodal and symmetric
  » Shape is Bell curve

→ Adheres to some guidelines about how variably the data is distributed around the mean




N(μ = 0, σ = 20)
N(μ = 219, σ = 70)
0
100
200
300

68 – 95 – 99.7% rule



Annie's Question

edureka!

A HR manager has consolidated the appraisal feedback obtained from various managers and has given ratings for a large number of employees at a corporate. If the HR manager reports the mean ratings as 60, the minimum as 30, and the maximum as 90, what could be the (approx.) standard deviation?

i. 1
ii. 10
iii. 30
iv. 45

Annie's Answer

edureka!

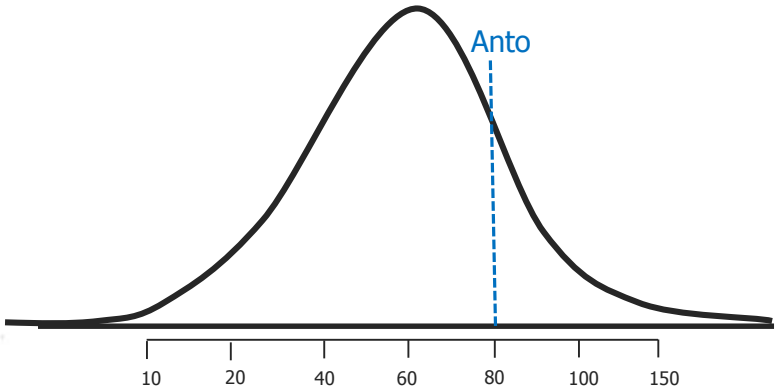
Ans. (ii) 10

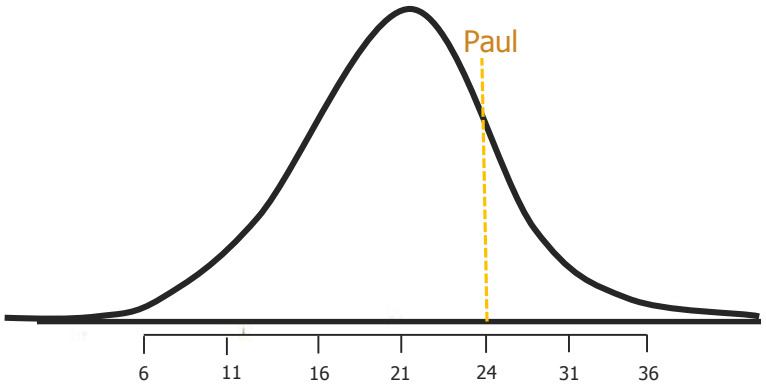
In India, different universities and colleges have different grading methods. Few students from across the country has applied for a software engineering position at a company.

The manager wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Anto, who earned an 80 on her Maharashtra University Exams, or Paul who scored a 24 on his University Exam at Bhopal?

$$Maharastra\ scores \approx N(Mean = 50,\ SD = 30)$$

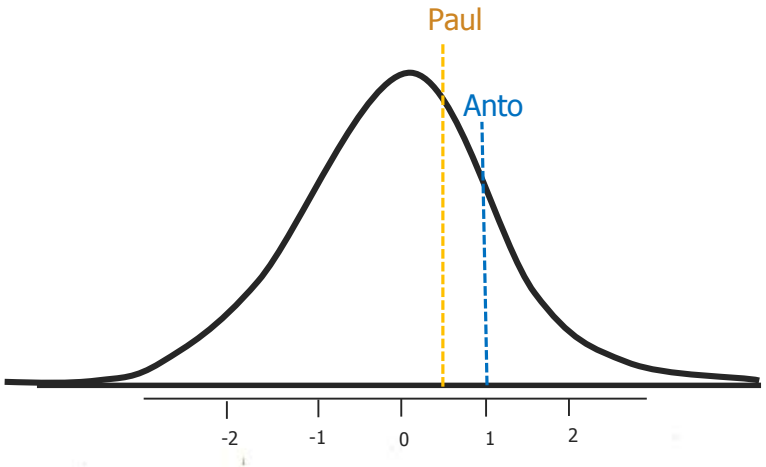
$$Bhopal\ scores \approx N(Mean = 21,\ SD = 5)$$





$$Anto: \frac{80 - 50}{30} = 1$$

$$Paul: \frac{24 - 21}{5} = 0.6$$



Standardizing with Z scores

→ number of standard deviations above or below the mean

→ Z score of mean = 0

→ Rare observations: |Z| > 2

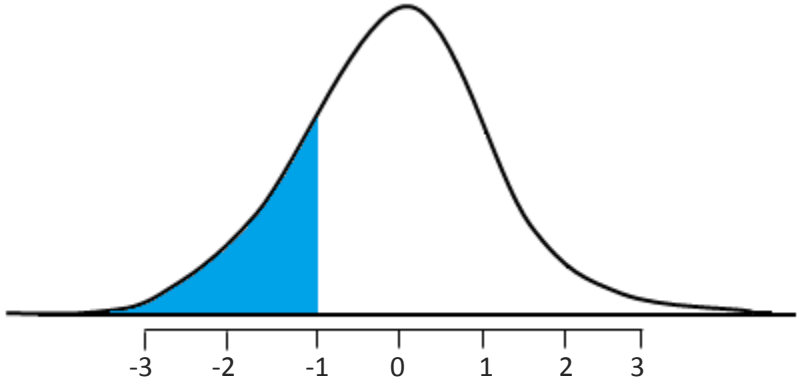
$$Z = \frac{Observation - Mean}{SD}$$

Percentiles

→ Percentile is the percentage of observations that fall below a given data point

→ The area below the probability distribution curve to the left of that observation

→ R command (pnorm(val, mean, sd))



Annie's Question

edureka!

Employees salary in a company are distributed normally with mean 50,000 Rs and standard deviation 30,000 Rs. Ram earns 70,000 Rs. How many percent of employees is Ram earn higher than their salaries?

Annie's Answer

edureka!

Ans. 74.75%

Annie's Question

A friend of you say that he is the top 5% of the salary provided in that company. What is the lowest salary that he could be getting (the same parameters in the previous problem apply here → mean : 50000, standard deviation = 30000)

Annie's Answer

edureka!

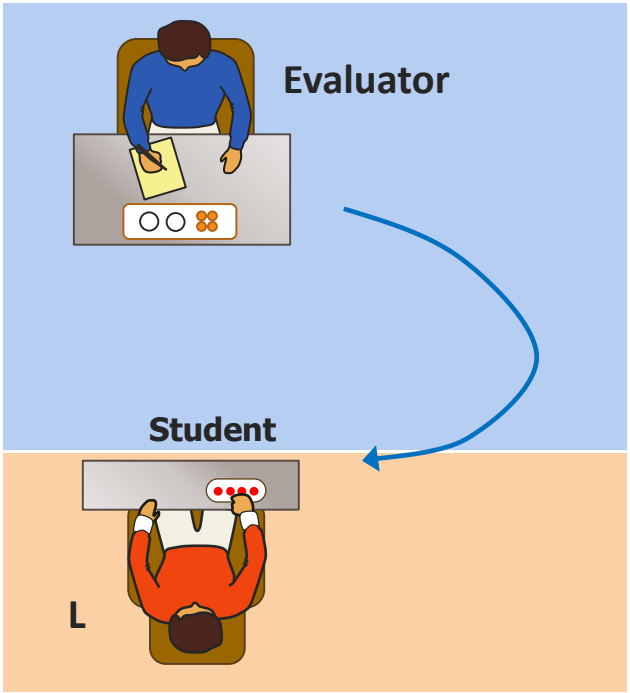
Ans. 88,446 Rs

Binomial Distribution

→ Definition, properties, conditions

→ Calculating probabilities

→ Mean and standard deviation



P(fail) = 0.65

Bernouilli Random Variables

→ Each student in the experiment can be thought of as a trial

→ A person is labeled a success if he/she passes the exam

→ Since only 35% of people pass the exam, probability of success is $p = 0.35$

→ When an individual trial has only two possible outcomes, it is called a Bernouilli random variable

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will pass the exam?

→ Four individuals
- A. Anto
- B. Binto
- C. Cinto
- D. Dinto

→ Multiple scenarios where "exactly 1 passes"

Scenario 1:

$$\frac{0.35}{(A)\ \text{pass}} \times \frac{0.65}{(B)\ fail} \times \frac{0.65}{(C)\ fail} \times \frac{0.65}{(D)\ fail} = 0.0961$$

OR

Scenario 2:

$$\frac{0.65}{(A)\ fail} \times \frac{0.35}{(B)\ \text{pass}} \times \frac{0.65}{(C)\ fail} \times \frac{0.65}{(D)\ fail} = 0.0961$$

OR

Scenario 3:

$$\frac{0.65}{(A)\ fail} \times \frac{0.65}{(B)\ fail} \times \frac{0.35}{(C)\ \text{fail}} \times \frac{0.65}{(D)\ pass} = 0.0961$$

OR

Scenario 4:

$$\frac{0.65}{(A)\ fail} \times \frac{0.65}{(B)\ fail} \times \frac{0.65}{(C)\ fail} \times \frac{0.35}{(D)\ \text{pass}} = 0.0961$$

+

4 X 0.0961 = 0.3844

Binomial Distribution

The binomial distribution describes the probability of having exactly k successes in n independent Bernouilli trials with probability of success *p*

# of scenarios X *P(single scenario)*

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

"n choose k"

$$p^k \left(1-p\right)^{(n-k)}$$

Probability of success to the power of number of successes

Probability of failure to the power of number of failures

How many scenarios yield 1 success in 5 trials?

$$n = 5, \quad k = 1$$

$$\binom{5}{1} = \frac{5!}{1! \times (5-1)!}$$

$$= \frac{5 \times 4 \times 3 \times 2 \times 1}{1 \times 4 \times 3 \times 2 \times 1} = 5$$

SSFFFFFFFF
SFSFFFFFFF
SFFSFFFFFF

How many scenarios yield 2 success in 10 trials?

$$n = 10, \quad k = 2$$

$$\binom{10}{2} = \frac{10!}{2! \times 8!}$$

$$= \frac{10 \times 9 \times 8!}{2 \times 1 \times 8!} = 45$$

Binomial Distribution:

If *p* represents probability of success, (1-*p*) represents probability of failure, *n* represents number of independent trials, and *k* represents number of successes.

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k \left(1 - p\right)^{(n-k)}$$

$$\text{where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Conditions

1. The trials must be independent.
2. The number of trials, n, must be fixed.
3. Each trial outcome must be classified as a success or a failure.

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will pass the exam?

$$n = 10$$
$$p = 0.24$$
$$1 - p = 0.76$$
$$k = 8$$

$$P(k = 8) = \binom{10}{8} 0.24^8 \times 0.76^2$$
$$= \frac{10!}{8! \times 2!} 0.24^8 \times 0.76^2$$
$$= 0.0002861078$$

dbinom(k, size, p)

Among a random sample of 100 women, how many would you expect to going work? p =0.24

$$\mu = 100 \times 0.24 = 24$$

$$Expected\ value\ (mean)\ of\ binomial\ distribution: \mu = np$$

$$S\tan dard\ deviation\ of\ binomial\ distribution: \sigma = \sqrt{np(1-p)}$$

$$\sigma = \sqrt{100 \times 0.24 \times 0.76} = 4.27$$

Project – Part 3

Variable Descriptions in the Data

edureka!

In order to understand the data, one has to follow the following variable descriptions:

Serial No	Variable	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number

Variable Descriptions in the Data (Contd.)

edureka!

Serial No	Variable	Description
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)

Variable Descriptions in the Data (Contd.)

edureka!

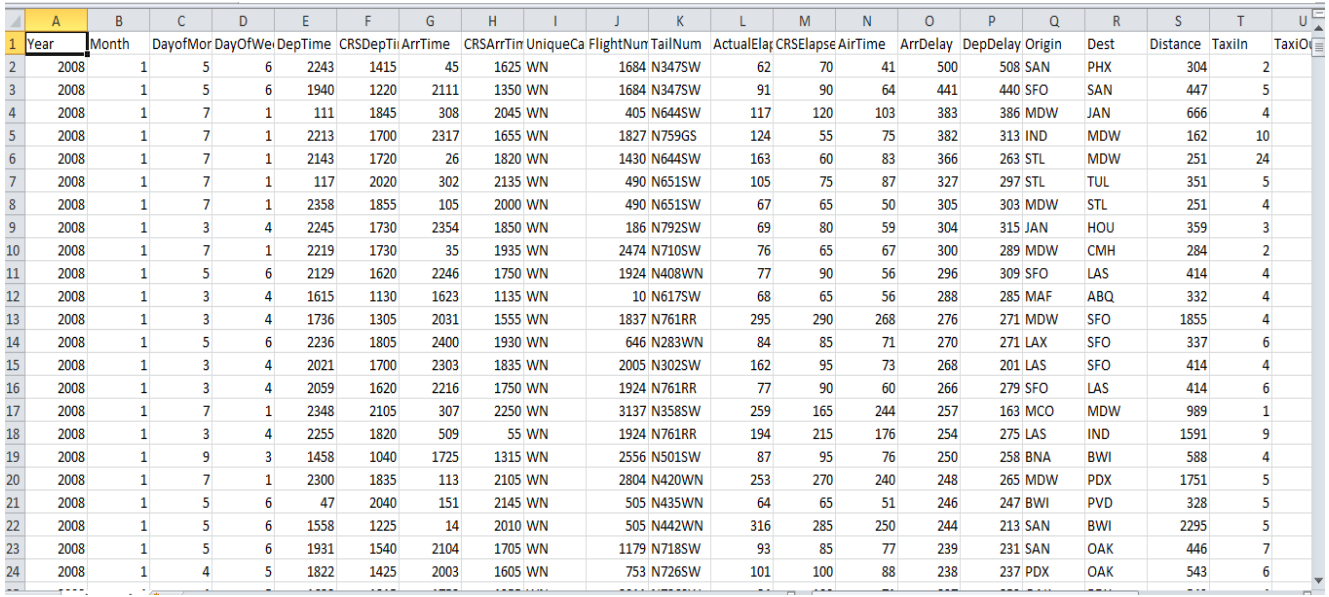
Serial No	Variable	Description
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

Snapshot of the Dataset

edureka!

You can take any of the years and try to solve the following problems.

A screenshot containing the 25 first lines may look like this:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Year	Month	DayofMor	DayOfWe	DepTime	CRSDepTi	ArrTime	CRSArrTin	UniqueCa	FlightNun	TailNum	ActualEla	CRSElapse	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiO
2	2008	1	5	6	2243	1415	45	1625	WN	1684	N347SW	62	70	41	500	508	SAN	PHX	304	2	
3	2008	1	5	6	1940	1220	2111	1350	WN	1684	N347SW	91	90	64	441	440	SFO	SAN	447	5	
4	2008	1	7	1	111	1845	308	2045	WN	405	N644SW	117	120	103	383	386	MDW	JAN	666	4	
5	2008	1	7	1	2213	1700	2317	1655	WN	1827	N759GS	124	55	75	382	313	IND	MDW	162	10	
6	2008	1	7	1	2143	1720	26	1820	WN	1430	N644SW	163	60	83	366	263	STL	MDW	251	24	
7	2008	1	7	1	117	2020	302	2135	WN	490	N651SW	105	75	87	327	297	STL	TUL	351	5	
8	2008	1	7	1	2358	1855	105	2000	WN	490	N651SW	67	65	50	305	303	MDW	STL	251	4	
9	2008	1	3	4	2245	1730	2354	1850	WN	186	N792SW	69	80	59	304	315	JAN	HOU	359	3	
10	2008	1	7	1	2219	1730	35	1935	WN	2474	N710SW	76	65	67	300	289	MDW	CMH	284	2	
11	2008	1	5	6	2129	1620	2246	1750	WN	1924	N408WN	77	90	56	296	309	SFO	LAS	414	4	
12	2008	1	3	4	1615	1130	1623	1135	WN	10	N617SW	68	65	56	288	285	MAF	ABQ	332	4	
13	2008	1	3	4	1736	1305	2031	1555	WN	1837	N761RR	295	290	268	276	271	MDW	SFO	1855	4	
14	2008	1	5	6	2236	1805	2400	1930	WN	646	N283WN	84	85	71	270	271	LAX	SFO	337	6	
15	2008	1	3	4	2021	1700	2303	1835	WN	2005	N302SW	162	95	73	268	201	LAS	SFO	414	4	
16	2008	1	3	4	2059	1620	2216	1750	WN	1924	N761RR	77	90	60	266	279	SFO	LAS	414	6	
17	2008	1	7	1	2348	2105	307	2250	WN	3137	N358SW	259	165	244	257	163	MCO	MDW	989	1	
18	2008	1	3	4	2255	1820	509	55	WN	1924	N761RR	194	215	176	254	275	LAS	IND	1591	9	
19	2008	1	9	3	1458	1040	1725	1315	WN	2556	N501SW	87	95	76	250	258	BNA	BWI	588	4	
20	2008	1	7	1	2300	1835	113	2105	WN	2804	N420WN	253	270	240	248	265	MDW	PDX	1751	5	
21	2008	1	5	6	47	2040	151	2145	WN	505	N435WN	64	65	51	246	247	BWI	PVD	328	5	
22	2008	1	5	6	1558	1225	14	2010	WN	505	N442WN	316	285	250	244	213	SAN	BWI	2295	5	
23	2008	1	5	6	1931	1540	2104	1705	WN	1179	N718SW	93	85	77	239	231	SAN	OAK	446	7	
24	2008	1	4	5	1822	1425	2003	1605	WN	753	N726SW	101	100	88	238	237	PDX	OAK	543	6	

Module-3 Problem Statement

edureka!

1. Suppose arrival delays of flights belonging to "AA" are normally distributed with mean 15 minutes and standard deviation 3 minutes. If the "AA" plans to announce a scheme where it will give 50% cash back if their flights are delayed by 20 minutes, how much percentage of the trips "AA" is supposed to loose this money. (Hint: pnorm)

2. Assume that 65% of flights are diverted due to bad weather through the Weather System. What is the probability that in a random sample of 10 flights, 6 are diverted through the Weather System. (Hint: dbinorm)

3. Do linear regression between the Arrival Delay and Departure Delay of the flights.

4. Find out the confidence interval of the fitted linear regression line.

5. Perform a multiple linear regression between the Arrival Delay along with the Departure Delay and Distance travelled by flights.

QUESTIONS

Survey

edureka!

Your feedback is important to us, be it a compliment, a suggestion or a complaint. It helps us to make the course better!

Please spare few minutes to take the survey after the webinar.

edureka!