```
import pandas as pd

df_uber = pd.read_csv(r'C:\Users\LENOVO\Downloads\archive (1)\
ncr_ride_bookings.csv')
df_uber.head(15)
```

```
          Date      Time    Booking ID       Booking Status   Customer
ID  \
0   2024-03-23  12:29:38  "CNR5884300"      No Driver Found
"CID1982111"
1   2024-11-29  18:01:39  "CNR1326809"           Incomplete
"CID4604802"
2   2024-08-23  08:56:10  "CNR8494506"            Completed
"CID9202816"
3   2024-10-21  17:17:25  "CNR8906825"            Completed
"CID2610914"
4   2024-09-16  22:08:00  "CNR1950162"            Completed
"CID9933542"
5   2024-02-06  09:44:56  "CNR4096693"            Completed
"CID4670564"
6   2024-06-17  15:45:58  "CNR2002539"            Completed
"CID6800553"
7   2024-03-19  17:37:37  "CNR6568000"            Completed
"CID8610436"
8   2024-09-14  12:49:09  "CNR4510807"      No Driver Found
"CID7873618"
9   2024-12-16  19:06:48  "CNR7721892"           Incomplete
"CID5214275"
10  2024-06-14  16:24:12  "CNR9070334"            Completed
"CID6680340"
11  2024-09-18  08:09:38  "CNR9551927"      No Driver Found
"CID7568143"
12  2024-06-25  22:44:15  "CNR4386945"  Cancelled by Driver
"CID5543520"
13  2024-09-11  19:29:39  "CNR2987763"            Completed
"CID2669710"
14  2024-10-18  18:28:53  "CNR8962232"            Completed
"CID1789354"

     Vehicle Type      Pickup Location      Drop Location  Avg VTAT
Avg CTAT  \
0           eBike          Palam Vihar            Jhilmil       NaN
NaN
1        Go Sedan        Shastri Nagar  Gurgaon Sector 56       4.9
14.0
2            Auto              Khandsa      Malviya Nagar      13.4
25.8
3   Premier Sedan  Central Secretariat           Inderlok      13.1
28.5
4            Bike     Ghitorni Village        Khan Market       5.3
```

```
19.6
5            Auto                AIIMS        Narsinghpur       5.1
18.1
6         Go Mini             Vaishali       Punjabi Bagh       7.1
20.4
7            Auto          Mayur Vihar          Cyber Hub      12.1
16.5
8        Go Sedan      Noida Sector 62    Noida Sector 18       NaN
NaN
9            Auto               Rohini       Adarsh Nagar       6.1
26.0
10           Auto         Udyog Bhawan   Dwarka Sector 21       7.7
18.9
11           Auto         Vidhan Sabha              AIIMS       NaN
NaN
12          eBike          Patel Chowk  Kherki Daula Toll       4.6
NaN
13        Go Mini        Malviya Nagar   Ghitorni Village      12.2
28.2
14        Go Mini              Madipur          GTB Nagar      14.0
30.9

    ...  Reason for cancelling by Customer Cancelled Rides by
Driver  \
0   ...                                NaN                       NaN

1   ...                                NaN                       NaN

2   ...                                NaN                       NaN

3   ...                                NaN                       NaN

4   ...                                NaN                       NaN

5   ...                                NaN                       NaN

6   ...                                NaN                       NaN

7   ...                                NaN                       NaN

8   ...                                NaN                       NaN

9   ...                                NaN                       NaN

10  ...                                NaN                       NaN

11  ...                                NaN                       NaN

12  ...                                NaN                       1.0

13  ...                                NaN                       NaN
```

```
14  ...                                NaN                       NaN


       Driver Cancellation Reason Incomplete Rides  Incomplete Rides
Reason  \
0                             NaN              NaN
NaN
1                             NaN              1.0        Vehicle
Breakdown
2                             NaN              NaN
NaN
3                             NaN              NaN
NaN
4                             NaN              NaN
NaN
5                             NaN              NaN
NaN
6                             NaN              NaN
NaN
7                             NaN              NaN
NaN
8                             NaN              NaN
NaN
9                             NaN              1.0              Other
Issue
10                            NaN              NaN
NaN
11                            NaN              NaN
NaN
12  Personal & Car related issues              NaN
NaN
13                            NaN              NaN
NaN
14                            NaN              NaN
NaN

   Booking Value  Ride Distance  Driver Ratings  Customer Rating  \
0            NaN            NaN             NaN              NaN
1          237.0           5.73             NaN              NaN
2          627.0          13.58             4.9              4.9
3          416.0          34.02             4.6              5.0
4          737.0          48.21             4.1              4.3
5          316.0           4.85             4.1              4.6
6          640.0          41.24             4.0              4.1
7          136.0           6.56             4.4              4.2
8            NaN            NaN             NaN              NaN
9          135.0          10.36             NaN              NaN
10         181.0          19.84             4.2              4.9
11           NaN            NaN             NaN              NaN
```

```
12           NaN            NaN             NaN              NaN
13         394.0          21.44             4.1              4.7
14         836.0          39.55             4.7              4.4

    Payment Method
0              NaN
1              UPI
2       Debit Card
3              UPI
4              UPI
5              UPI
6              UPI
7              UPI
8              NaN
9             Cash
10            Cash
11             NaN
12             NaN
13             UPI
14             UPI

[15 rows x 21 columns]
```

```
df_uber[['Vehicle Type', 'Booking Value']]
```

```
         Vehicle Type  Booking Value
0               eBike            NaN
1            Go Sedan          237.0
2                Auto          627.0
3       Premier Sedan          416.0
4                Bike          737.0
...               ...            ...
149995        Go Mini          475.0
149996        Go Mini         1093.0
149997       Go Sedan          852.0
149998           Auto          333.0
149999  Premier Sedan          806.0

[150000 rows x 2 columns]
```

```
df_clear = df_uber[['Vehicle Type', 'Booking Value']].dropna()

display(df_clear)
```

```
         Vehicle Type  Booking Value
1            Go Sedan          237.0
2                Auto          627.0
3       Premier Sedan          416.0
4                Bike          737.0
5                Auto          316.0
...               ...            ...
```

```
149995        Go Mini          475.0
149996        Go Mini         1093.0
149997       Go Sedan          852.0
149998           Auto          333.0
149999  Premier Sedan          806.0

[102000 rows x 2 columns]
```

```python
df_clear.isnull().any(axis=1)
```

```
1         False
2         False
3         False
4         False
5         False
          ...  
149995    False
149996    False
149997    False
149998    False
149999    False
Length: 102000, dtype: bool
```

```python
## Making two grouos Auto and Bike to check if the mean is sambe
between them or not
#Null Hypothesis is mean of Auto = Bike
# Alternative Hypothesis is mean Auto != Bike
Auto = df_clear[df_clear['Vehicle Type'] == 'Auto']['Booking Value']
Bike = df_clear[df_clear['Vehicle Type'] == 'Bike']['Booking Value']

print(Auto)
print(Bike)
```

```
2         627.0
5         316.0
7         136.0
9         135.0
10        181.0
          ...  
149964    643.0
149969    524.0
149989     75.0
149991    597.0
149998    333.0
Name: Booking Value, Length: 25415, dtype: float64
4         737.0
28        304.0
47        453.0
74        633.0
82        224.0
```

```
          ...
149962    227.0
149967    194.0
149975    507.0
149985    193.0
149988     96.0
Name: Booking Value, Length: 15362, dtype: float64
```

```python
from scipy import stats

t_stats, p_value = stats.ttest_ind(Auto, Bike, equal_var=False)

print(f't_stats is {t_stats} and p_value is {p_value}')
```

```
t_stats is -0.8560975357459055 and p_value is 0.3919502919192158
```

Since the p-value (0.392) > 0.05, we fail to reject the null hypothesis. This means there is no statistically significant difference in the average Booking Value between Auto and Bike rides.

Two-sample independent t-test

Needs 1 numerical column (the measurement)

Needs 1 categorical column (to split into 2 groups, like Auto vs Bike)

`Example`

`Numerical = Booking Value`

`Categorical = Vehicle Type (Auto vs Bike)`

Paired-sample t-test

Needs 2 numerical columns

Both measured on the same row / same subject / same ride

No categorical grouping needed

We look at the difference between the two columns for each row

`Example`

`Numerical columns = Driver Ratings and Customer Rating`

```
Each row = one ride, with two scores
```

```
## Running T test paired Sample
df_uber
```

```
              Date      Time    Booking ID   Booking Status   Customer
ID  \
0       2024-03-23  12:29:38  "CNR5884300"  No Driver Found
"CID1982111"
1       2024-11-29  18:01:39  "CNR1326809"       Incomplete
"CID4604802"
2       2024-08-23  08:56:10  "CNR8494506"        Completed
"CID9202816"
3       2024-10-21  17:17:25  "CNR8906825"        Completed
"CID2610914"
4       2024-09-16  22:08:00  "CNR1950162"        Completed
"CID9933542"
...            ...       ...           ...              ...
...
149995  2024-11-11  19:34:01  "CNR6500631"        Completed
"CID4337371"
149996  2024-11-24  15:55:09  "CNR2468611"        Completed
"CID2325623"
149997  2024-09-18  10:55:15  "CNR6358306"        Completed
"CID9925486"
149998  2024-10-05  07:53:34  "CNR3030099"        Completed
"CID9415487"
149999  2024-03-10  15:38:03  "CNR3447390"        Completed
"CID4108667"

         Vehicle Type         Pickup Location      Drop Location  Avg
VTAT  \
0               eBike             Palam Vihar            Jhilmil
NaN
1            Go Sedan           Shastri Nagar  Gurgaon Sector 56
4.9
2                Auto                 Khandsa      Malviya Nagar
13.4
3       Premier Sedan     Central Secretariat           Inderlok
13.1
4                Bike        Ghitorni Village        Khan Market
5.3
...               ...                     ...                ...
...
149995        Go Mini                 MG Road           Ghitorni
10.2
149996        Go Mini        Golf Course Road         Akshardham
5.1
149997       Go Sedan  Satguru Ram Singh Marg           Jor Bagh
2.7
```

```
149998           Auto               Ghaziabad         Saidulajab
6.9
149999  Premier Sedan         Ashok Park Main  Gurgaon Sector 29
3.5

        Avg CTAT  ...  Reason for cancelling by Customer  \
0            NaN  ...                                NaN
1           14.0  ...                                NaN
2           25.8  ...                                NaN
3           28.5  ...                                NaN
4           19.6  ...                                NaN
...          ...  ...                                ...
149995      44.4  ...                                NaN
149996      30.8  ...                                NaN
149997      23.4  ...                                NaN
149998      39.6  ...                                NaN
149999      33.7  ...                                NaN

       Cancelled Rides by Driver  Driver Cancellation Reason
Incomplete Rides  \
0                            NaN                         NaN
NaN
1                            NaN                         NaN
1.0
2                            NaN                         NaN
NaN
3                            NaN                         NaN
NaN
4                            NaN                         NaN
NaN
...                          ...                         ...
...
149995                       NaN                         NaN
NaN
149996                       NaN                         NaN
NaN
149997                       NaN                         NaN
NaN
149998                       NaN                         NaN
NaN
149999                       NaN                         NaN
NaN

        Incomplete Rides Reason Booking Value  Ride Distance  Driver
Ratings  \
0                           NaN           NaN            NaN
NaN
1             Vehicle Breakdown         237.0           5.73
NaN
2                           NaN         627.0          13.58
```

```
4.9
3                           NaN          416.0          34.02
4.6
4                           NaN          737.0          48.21
4.1
...                         ...            ...            ...
...
149995                      NaN          475.0          40.08
3.7
149996                      NaN         1093.0          21.31
4.8
149997                      NaN          852.0          15.93
3.9
149998                      NaN          333.0          45.54
4.1
149999                      NaN          806.0          21.19
4.6

        Customer Rating  Payment Method
0                   NaN             NaN
1                   NaN             UPI
2                   4.9      Debit Card
3                   5.0             UPI
4                   4.3             UPI
...                 ...             ...
149995              4.1     Uber Wallet
149996              5.0             UPI
149997              4.4            Cash
149998              3.7             UPI
149999              4.9     Credit Card

[150000 rows x 21 columns]
```

```
clean_data2 = df_uber[['Driver Ratings','Customer Rating']].dropna()
```

```
clean_data2
```

```
        Driver Ratings  Customer Rating
2                  4.9              4.9
3                  4.6              5.0
4                  4.1              4.3
5                  4.1              4.6
6                  4.0              4.1
...                ...              ...
149995             3.7              4.1
149996             4.8              5.0
149997             3.9              4.4
149998             4.1              3.7
149999             4.6              4.9
```

```
[93000 rows x 2 columns]

## Taking two numerical columns.
# Null hypothesis is mean of Driver Ratings = Customer Rating
# alternative hypothesis is mean of Driver Ratings != Customer Rating
Driver_Ratings = df_uber['Driver Ratings']
Customer_Rating = df_uber['Customer Rating']

Driver_Ratings = Driver_Ratings.dropna()

Customer_Rating = Customer_Rating.dropna()

t_stat, p_value = stats.ttest_rel(Driver_Ratings, Customer_Rating)

print(f't_stat is {t_stat} and p_value is {p_value}')

t_stat is -85.5481993199033 and p_value is 0.0
```

We performed a paired sample t-test to compare Driver Ratings and Customer Ratings for the same rides. The test gave t = -85.55 and p < 0.001, so we reject the null hypothesis. This shows that there is a significant difference between how drivers and customers rate rides, with drivers tending to give lower ratings on average

T_Stats Meaning

One sample I have done in my previous project

`1. One-Sample t-test`

Is my sample mean different from some assumed value?

Example:

Boss assumes average Booking Value = 1000

We take a random sample and test.

t_stat meaning:

Large positive → sample mean is much greater than 1000.

Large negative → sample mean is much less than 1000.

Close to 0 → sample mean is about the same as 1000.

`2. Two-Sample t-test (independent groups)`

Are the averages of two groups different?

Example:

Compare Auto vs Bike Booking Values.

t_stat meaning:

Large positive → Auto mean is greater than Bike mean.

Large negative → Auto mean is less than Bike mean.

Close to 0 → Auto and Bike means are similar.

`3. Paired-Sample t-test (dependent samples)`

Are two measurements on the same ride different?

Example:

Compare Driver Ratings vs Customer Ratings for the same trip.

t_stat meaning:

Large positive → Driver Ratings are higher than Customer Ratings.

Large negative → Driver Ratings are lower than Customer Ratings.

Close to 0 → Both ratings are similar.

Z Test

Lets run a Z test on this data. Two Sample

```
Cleaned_Data = df_uber[['Pickup Location', 'Booking Value']].dropna()

Cleaned_Data
```

```
             Pickup Location  Booking Value
1              Shastri Nagar          237.0
2                    Khandsa          627.0
3       Central Secretariat          416.0
4           Ghitorni Village          737.0
5                      AIIMS          316.0
```

```
...                          ...           ...
149995                   MG Road         475.0
149996          Golf Course Road        1093.0
149997  Satguru Ram Singh Marg         852.0
149998                 Ghaziabad         333.0
149999           Ashok Park Main         806.0

[102000 rows x 2 columns]
```

```python
##Grouping Data
Shastri_Nagar = Cleaned_Data[Cleaned_Data['Pickup Location'] ==
'Shastri Nagar']['Booking Value']
Khandsa = Cleaned_Data[Cleaned_Data['Pickup Location'] == 'Khandsa']
['Booking Value']


from statsmodels.stats.weightstats import ztest

z_stat, p_value = ztest(Shastri_Nagar, Khandsa, alternative= 'two-
sided')

print(f'z_stat is {z_stat} and p_value is {p_value}')
```

```
z_stat is 0.6250622877824882 and p_value is 0.5319301780550996
```

We compared the mean Booking Values between Shastri Nagar and Khandsa using a two-sample Z-test. Since p-value > 0.05, we fail to reject the null hypothesis. This means the average booking values are statistically the same between the two pickup locations

T-test vs Z-test (Simple)

Both are used to test means (1-sample, 2-sample, or paired).
T-test = when sample size is small **Z-test** = when sample size is large.