ANÁLISE DE VIAGENS POR APLICATIVO

SERÁ REALIZADA A LEITURA,CORREÇÃO E ANÁLISE DOS CONJUNTOS DE DADOS DAS VIAGENS POR APLICATIVO NA CIDADE DE CHICAGO. COM OS DADOS LIMPOS, IREMOS EXTRAIR DOS DADOS OS PRINCIPAIS DESTINOS E AS PRINCIPAIS EMPRESAS DE TÁXI DA CIDADE. SERÁ QUE NOS DIAS DE SÁBADOS CHUVOSOS A DURAÇÃO MÉDIA DOS PASSEIOS SOFRE ALTERAÇÃO? VAMOS VERIFICAR A HIPÓTESE!

INICIALIZAÇÃO

In [3]:

```
#CARREGANDO AS BIBLIOTECAS
import pandas as pd
from scipy.stats import levene
from scipy import stats as st
import numpy as np
from math import factorial
import matplotlib.pyplot as plt
```

CARREGANDO OS DADOS

In [4]:

```
#NÚMERO DE CORRIDAS POR CADA EMPRESA ENTRE 15 E 16 DE NOVEMBRO DE 2017
corridas_por_empresa= pd.read_csv('/datasets/project_sql_result_01.csv')

# NÚMERO MÉDIO DE VIAGENS TERMINADA EM CADA BAIRRO DE CHICAGO EM NOVEMBRO DE 2017
corridas_por_bairro= pd.read_csv('/datasets/project_sql_result_04.csv')

# DADOS DAS VIAGENS DO LOOP PARA O AEROPORTO INTERNACIONAL O'HARE
loop_aeroporto= pd.read_csv('/datasets/project_sql_result_07.csv')
```

NÚMERO DE CORRIDAS POR CADA EMPRESA ENTRE 15 E 16 DE NOVEMBRO DE 2017

In [5]:

```
corridas_por_empresa
```

Out[5]:

	company_name	trips_amount
0	Flash Cab	19558
1	Taxi Affiliation Services	11422
2	Medallion Leasing	10367
3	Yellow Cab	9888
4	Taxi Affiliation Service Yellow	9299
...	...	...
59	4053 - 40193 Adwar H. Nikola	7
60	2733 - 74600 Benny Jona	7
61	5874 - 73628 Sergey Cab Corp.	5
62	2241 - 44667 - Felman Corp	3
63	3556 - 36214 RC Andrews Cab	2

64 rows × 2 columns

In [6]:

```
corridas_por_empresa.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 2 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   company_name  64 non-null     object
 1   trips_amount  64 non-null     int64 
dtypes: int64(1), object(1)
memory usage: 1.1+ KB
```

VERIFICANDO VALORES DUPLICADOS

In [8]:

```
corridas_por_empresa.duplicated().sum()
```

Out[8]:

```
0
```

In [9]:

```
corridas_por_empresa.describe()
```

Out[9]:

	trips_amount
count	64.000000
mean	2145.484375
std	3812.310186
min	2.000000
25%	20.750000
50%	178.500000
75%	2106.500000
max	19558.000000

As 10 Principais Empresas em Números de Viagens

In [218]:

```
corridas_por_empresa.sort_values(by='trips_amount', ascending=False).head(10)
```

Out[218]:

	company_name	trips_amount
0	Flash Cab	19558
1	Taxi Affiliation Services	11422
2	Medallion Leasing	10367
3	Yellow Cab	9888
4	Taxi Affiliation Service Yellow	9299
5	Chicago Carriage Cab Corp	9181
6	City Service	8448
7	Sun Taxi	7701
8	Star North Management LLC	7455
9	Blue Ribbon Taxi Association Inc.	5953

Conclusão do Número de Corridas por Empresa

Os tipos das colunas estão corretos e não possuímos valores ausentes.

NÚMERO MÉDIO DE VIAGENS TERMINADA EM CADA BAIRRO DE CHICAGO EM NOVEMBRO DE 2017

In [219]:

```
corridas_por_bairro
```

Out[219]:

	dropoff_location_name	average_trips
0	Loop	10727.466667
1	River North	9523.666667
2	Streeterville	6664.666667
3	West Loop	5163.666667
4	O'Hare	2546.900000
...	...	...
89	Mount Greenwood	3.137931
90	Hegewisch	3.117647
91	Burnside	2.333333
92	East Side	1.961538
93	Riverdale	1.800000

94 rows × 2 columns

In [220]:

```
corridas_por_bairro.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 94 entries, 0 to 93
Data columns (total 2 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   dropoff_location_name  94 non-null     object 
 1   average_trips          94 non-null     float64
dtypes: float64(1), object(1)
memory usage: 1.6+ KB
```

VERIFICANDO VALORES DUPLICADOS

In [10]:

```
corridas_por_bairro.duplicated().sum()
```

Out[10]:

```
0
```

In [11]:

```
corridas_por_bairro.describe()
```

Out[11]:

	average_trips
count	94.000000
mean	599.953728
std	1714.591098
min	1.800000
25%	14.266667
50%	52.016667
75%	298.858333
max	10727.466667

Os 10 principais destinos das viagens

In [221]:

```
corridas_por_bairro_top10= corridas_por_bairro.sort_values(by='average_trips', ascending=
corridas_por_bairro_top10
```

Out[221]:

	dropoff_location_name	average_trips
0	Loop	10727.466667
1	River North	9523.666667
2	Streeterville	6664.666667
3	West Loop	5163.666667
4	O'Hare	2546.900000
5	Lake View	2420.966667
6	Grant Park	2068.533333
7	Museum Campus	1510.000000
8	Gold Coast	1364.233333
9	Sheffield & DePaul	1259.766667

conclusão do Número de Viagens por Bairro

Os tipos das colunas estão corretos e não possuímos valores ausentes.

DADOS DAS VIAGENS DO LOOP PARA O AEROPORTO INTERNACIONAL O'HARE

In [222]:

```
loop_aeroporto
```

Out[222]:

	start_ts	weather_conditions	duration_seconds
0	2017-11-25 16:00:00	Good	2410.0
1	2017-11-25 14:00:00	Good	1920.0
2	2017-11-25 12:00:00	Good	1543.0
3	2017-11-04 10:00:00	Good	2512.0
4	2017-11-11 07:00:00	Good	1440.0
...	...	...	...
1063	2017-11-25 11:00:00	Good	0.0
1064	2017-11-11 10:00:00	Good	1318.0
1065	2017-11-11 13:00:00	Good	2100.0
1066	2017-11-11 08:00:00	Good	1380.0
1067	2017-11-04 16:00:00	Bad	2834.0

1068 rows × 3 columns

In [223]:

```
loop_aeroporto.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1068 entries, 0 to 1067
Data columns (total 3 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   start_ts            1068 non-null   object 
 1   weather_conditions  1068 non-null   object 
 2   duration_seconds    1068 non-null   float64
dtypes: float64(1), object(2)
memory usage: 25.2+ KB
```

Conclusão Das viagens do Loop para o Aeroporto

Os dados não possuem valores ausentes. Entretanto, o tipo da coluna start_ts está incorreto, vamos corrigir.

In [224]:

```
# Transformando em DateTime
loop_aeroporto['start_ts']= pd.to_datetime(loop_aeroporto['start_ts'],format='%Y-%m-%d %H
```

In [225]:

```
# Enriquecendo os dados
loop_aeroporto['hours']= pd.DatetimeIndex(loop_aeroporto['start_ts']).hour
loop_aeroporto['day']= pd.DatetimeIndex(loop_aeroporto['start_ts']).day
```

In [226]:

```
loop_aeroporto
```

Out[226]:

	start_ts	weather_conditions	duration_seconds	hours	day
0	2017-11-25 16:00:00	Good	2410.0	16	25
1	2017-11-25 14:00:00	Good	1920.0	14	25
2	2017-11-25 12:00:00	Good	1543.0	12	25
3	2017-11-04 10:00:00	Good	2512.0	10	4
4	2017-11-11 07:00:00	Good	1440.0	7	11
...	...	...	...	...	...
1063	2017-11-25 11:00:00	Good	0.0	11	25
1064	2017-11-11 10:00:00	Good	1318.0	10	11
1065	2017-11-11 13:00:00	Good	2100.0	13	11
1066	2017-11-11 08:00:00	Good	1380.0	8	11
1067	2017-11-04 16:00:00	Bad	2834.0	16	4

1068 rows × 5 columns

In [227]:

```
loop_aeroporto.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1068 entries, 0 to 1067
Data columns (total 5 columns):
 #   Column              Non-Null Count  Dtype         
---  ------              --------------  -----         
 0   start_ts            1068 non-null   datetime64[ns]
 1   weather_conditions  1068 non-null   object        
 2   duration_seconds    1068 non-null   float64       
 3   hours               1068 non-null   int64         
 4   day                 1068 non-null   int64         
dtypes: datetime64[ns](1), float64(1), int64(2), object(1)
memory usage: 41.8+ KB
```

CONSTRUÇÃO GRÁFICO DOS DADOS

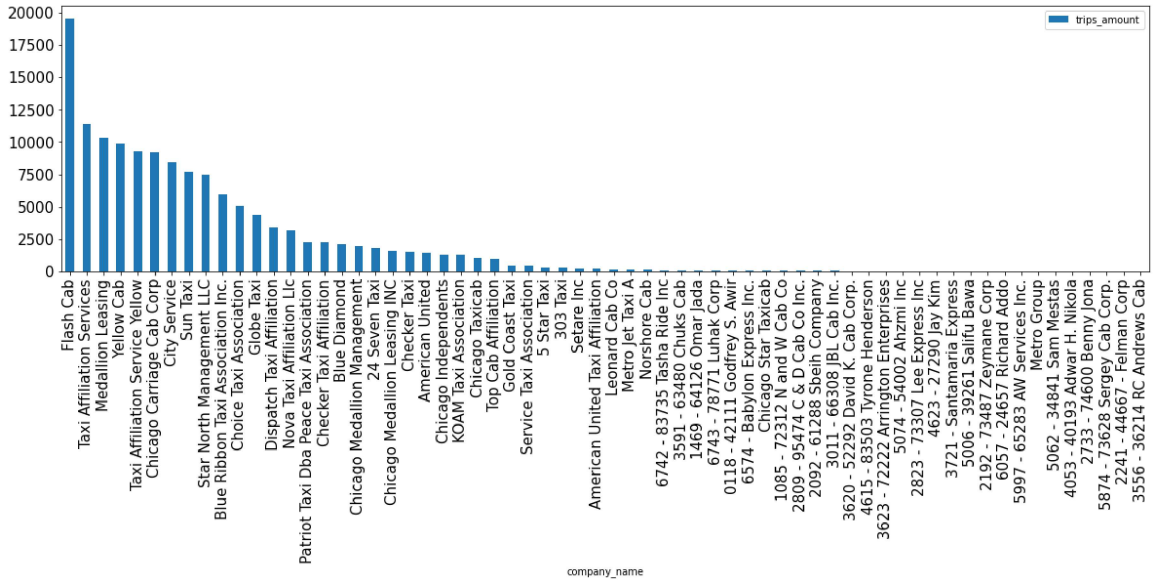
EMPRESAS DE TAXI E NÚMERO DE CORRIDAS

In [228]:

```
corridas_por_empresa.plot(x='company_name',y='trips_amount',kind='bar',figsize=(20,5),for
```

Out[228]:

```
<AxesSubplot:xlabel='company_name'>
```



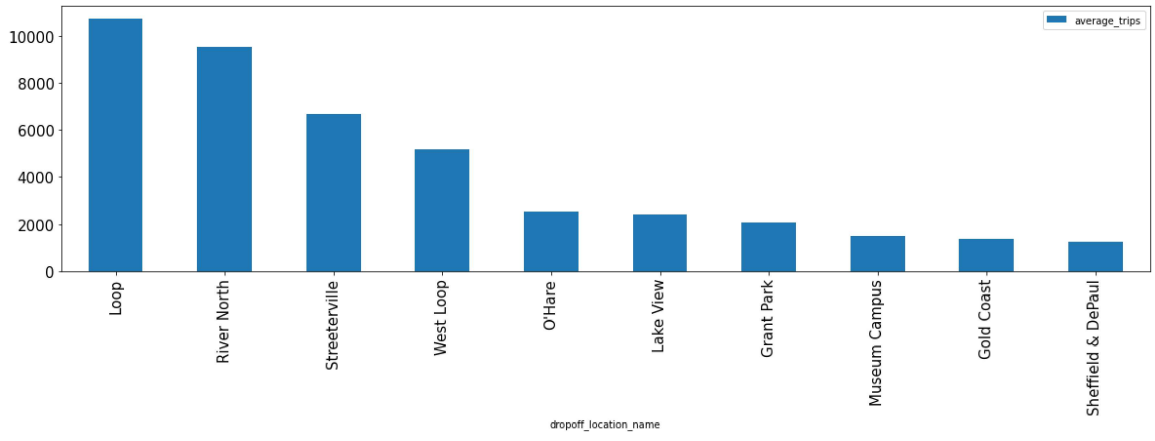
OS 10 PRINCIPAIS BAIRROS DE DESTINO E O NÚMERO DE CORRIDA

In [229]:

```
corridas_por_bairro_top10.plot(x='dropoff_location_name',y='average_trips',kind='bar',fig
```

Out[229]:

```
<AxesSubplot:xlabel='dropoff_location_name'>
```



Conclusão Dos Gráficos

Empresas Identificamos as pricipais empresas, em termos de número de corridas, da cidade de Chicago. A empresa Flash Cab é a destacada, pois possui um número de corridas muito superior aos números das outras empresas, quase duas vezes maior que a terceira empresa do ranking, a Medallion Leasing.

Bairros Identificado os 10 principais bairro de destino dos usuários, temos como destaque o Loop e o River North. Este primeiro, possui um número superior em destino ao bairro River North. Eles se destacam, pois a

TESTANDO HIPÓTESES

"A DURAÇÃO MÉDIA DOS PASSEIOS DO LOOP PARA O AEROPORTO INTERNACIONAL O'HARE MUDA NOS SÁBADOS CHUVOSOS."

In [233]:

```
# Aplicando teste de levene para definir o valor de equal_var
# H0; variâncias são iguais

sabado_chuvoso = loop_aeroporto.query('weather_conditions == "Bad"')
sabado_normal = loop_aeroporto.query('weather_conditions == "Good"')

sample_1= sabado_chuvoso['duration_seconds'][sabado_chuvoso['duration_seconds'].notna()].
sample_2= sabado_normal['duration_seconds'][sabado_normal['duration_seconds'].notna()].to
alpha= 0.05
stat, p = st.levene(sample_1,sample_2)
result_levene= p

if result_levene < alpha:
    print('Rejeita H0, variâncias são diferentes, equal_var = False para o teste')
else:
    print('Não rejeite H0, variâncias não diferem, equal_var = True para o teste')

var_sample= [np.var(x,ddof=1) for x in [sample_1,sample_2]]
print('Sample variances:', var_sample)
```

```
Não rejeite H0, variâncias não diferem, equal_var = True para o teste
Sample variances: [520294.086002483, 576382.009689509]
```

In [234]:

```
# TESTANDO A HIPÓTESE - "A DURAÇÃO MÉDIA DOS PASSEIOS DO LOOP PARA O AEROPORTO INTERNACIO

# HO; A DURAÇÃO MÉDIA DOS PASSEIOS DO LOOP PARA O AEROPORTO É IGUAL NOS SÁBADOS CHUVOSOS
# H1; A DURAÇÃO MÉDIA DOS PASSEIOS DO LOOP PARA O AEROPORTO INTERNACIONAL O'HARE MUDA NOS

alpha=0.05  #nível crítico de significância estatística
# se for menor que este valor as médias da classificação dos usuários são diferentes.

results= st.ttest_ind(loop_aeroporto[loop_aeroporto['weather_conditions']=='Bad']['durati
                      loop_aeroporto[loop_aeroporto['weather_conditions']=='Good']['durati
                      equal_var=True)

print('p-value:', results.pvalue)

if results.pvalue < alpha:
    print('A DURAÇÃO MÉDIA DOS PASSEIOS DO LOOP PARA O AEROPORTO INTERNACIONAL OHARE MUDA
else:
    print('NÃO PODEMOS REJEITAR A HIPÓTESE:A DURAÇÃO MÉDIA DOS PASSEIOS DO LOOP PARA O AE
```

```
p-value: 6.517970327099473e-12
A DURAÇÃO MÉDIA DOS PASSEIOS DO LOOP PARA O AEROPORTO INTERNACIONAL OHARE
MUDA NOS SÁBADOS CHUVOSOS
```

CONCLUSÃO GERAL

LEITURA E CORREÇÃO DOS DADOS

FOI REALIZADA A LEITURA DOS DATAFRAMES. ESPECIALMENTE NO DF REFERENTE AOS DADOS DAS VIAGENS DO LOOP PARA O AEROPORTO, MUDAMOS A TIPAGEM DE UMA DAS COLUNAS E O ENRIQUECEMOS COM MAIS DUAS COLUNAS, O DIA E A HORA.

Gráficos

UTILIZAMOS OS GRÁFICOS DE BARRAS E IDENTIFICAMOS OS 10 PRINCIPAIS DESTINOS, COM DESTAQUE PARA O LOOP E O RIVER NORTH. SENDO O DESTINO NO BAIRRO LOOP UM NÚMERO SUPERIOR AO BAIRRO RIVER NORTH.

IGUALMENTE, IDENTIFICAMOS AS EMPRESAS COM MAIORES NÚMEROS DE CORRIDAS DA CIDADE DE CHICAGO. A EMPRESA FLASH CAB É A QUE POSSUI MAIOR NÚMERO DE CORRIDA, MUITO SUPERIOR AOS NÚMEROS DAS OUTRAS EMPRESAS.

TESTANDO A HIPÓTESE:"A DURAÇÃO MÉDIA DOS PASSEIOS DO LOOP PARA O AEROPORTO INTERNACIONAL O'HARE MUDA NOS SÁBADOS CHUVOSOS."

PARA TESTAR A HIPÓTESE, FOI REALIZADO O TESTE DE LEVENE PARA AVALIAÇÃO DA VARIÂNCIA DOS DADOS PARA CONFIGURAR O PARÂMETRO 'EQUAL_VAR'. EM SEGUIDA, APLICA-SE O TESTE DE HIPÓTESE E CONFIRMAMOS A HIPÓTESE PRÉ ESTABELECIDA.

In [ ]: