title: "Untitled1.ipynb" format: pdf: toc: true number-sections: true colorlinks: true

Problem 1

(a) 1. There are missing values in the dataset. 2. Dates collected are not all in the same format. 3. Some of the values don't make sense (-999 for weight). 4. One of them doesn't have a family column. 5. All the plots seem to have the same type of information--it woul be better combined

(b) 1. Remove Nan values, question marks, etc. by going through the cells. 2. put the dates in standard format. 3. Remove those values. 4 and 5. Either throw away the data without a family column or make two plots--one with data with the family column and one without.

Problem 2

In [3]:
```
import pandas as pd
import matplotlib.pyplot as plt
import os
```

In [24]:
```
cities = ['amsterdam', 'athens', 'barcelona', 'berlin', 'budapest',
          'lisbon', 'london', 'paris', 'rome', 'vienna']
daytypes = ['weekdays', 'weekends']
datafolder = '/Users/AbigailLu/DATA_119 W26/Data'
alldata = []

for city in cities:
    for daytype in daytypes:
        filename = city + "_" + daytype + ".csv"
        filepath = os.path.join(datafolder, filename)
        airbnbs = pd.read_csv(filepath)

        airbnbs['city'] = city
        airbnbs['day_type'] = daytype
        alldata.append(airbnbs)

comdata = pd.concat(alldata, ignore_index=True)

print("Dimensions:" + str(comdata.shape))
```

```
#showing the price
weekendsdata = comdata[comdata['day_type'] == 'weekends']
meanweekendprice = weekendsdata['realSum'].mean()
print("Mean price for weekends in all cities: " + str(round(meanweekendprice, 2)))

weekdaysdata = comdata[comdata['day_type'] == 'weekdays']
meanweekdayprice = weekdaysdata['realSum'].mean()
print("Mean price for weekdays in all cities: " + str(round(meanweekdayprice, 2)))

#calculating difference
pricediff = meanweekendprice - meanweekdayprice
print("Difference: " + str(round(pricediff, 2)))
```

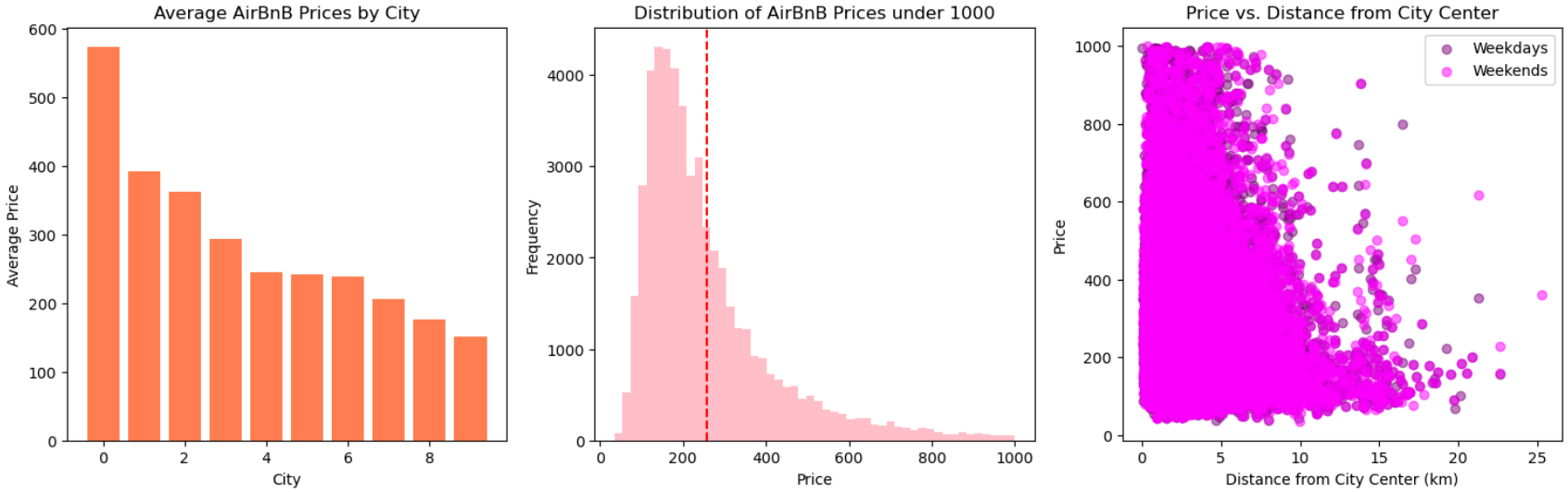
```
Dimensions:(51707, 22)
Mean price for weekends in all cities: 283.96
Mean price for weekdays in all cities: 275.68
Difference: 8.28
```

In [39]:

```
#graph 1
fig, axes = plt.subplots(1, 3, figsize=(18, 5))

cityavgprices = comdata.groupby('city')['realSum'].mean().sort_values(ascending=False)

axes[0].bar(range(len(city_avg_prices)), city_avg_prices.values, color = 'coral')
axes[0].set_xlabel('City')
axes[0].set_ylabel('Average Price')
axes[0].set_title('Average AirBnB Prices by City')

#graph 2
filteredprices = comdata[comdata['realSum'] < 1000]['realSum']

axes[1].hist(filteredprices, bins=50, color = 'pink')
axes[1].set_xlabel('Price')
axes[1].set_ylabel('Frequency')
axes[1].set_title('Distribution of AirBnB Prices under 1000')

meanprice = filteredprices.mean()
axes[1].axvline(meanprice, color='red', linestyle='--')

#graph 3
scatterdata = comdata[comdata['realSum'] < 1000]
weekdaysscat = scatterdata[scatterdata['day_type'] == 'weekdays']
```

```
axes[2].scatter(weekdaysscat['dist'], weekdaysscat['realSum'], color = 'purple', label = 'Weekdays', alpha
weekendsscat = scatterdata[scatterdata['day_type'] == 'weekends']
axes[2].scatter(weekendsscat['dist'], weekendsscat['realSum'], c = 'magenta', label = 'Weekends', alpha = (

axes[2].set_xlabel('Distance from City Center (km)')#, fontsize=11, fontweight='bold')
axes[2].set_ylabel('Price')#, fontsize=11, fontweight='bold')
axes[2].set_title('Price vs. Distance from City Center')#, fontsize=12, fontweight='bold')
axes[2].legend()
```

Out[39]: <matplotlib.legend.Legend at 0x11ad4f390>



In [40]:
```
caption_1 = """
Figure 1 - Bar Chart: Average AirBnB Prices by City
This bar chart displays the mean nightly rental price across ten European cities. This display gives us in
and numerical information (average price on y-axis). London shows the highest average price, followed by Pa
and Amsterdam, while Budapest and Athens offer the most affordable options.
"""

caption_2 = """
Figure 2 - Histogram: Distribution of AirBnB Prices under 1000
This histogram illustrates the frequency distribution of AirBnB listing prices under 1000 across all ten c
with most listings concentrated between 50-200 per night and as seen by the dashed line that doesn't line
the distribution of a single numerical variable.
"""
```

```
caption_3 = """
Figure 3 - Scatterplot: Price vs. Distance from City Center
This scatterplot examines the relationship between an AirBnB's distance from the city center (in kilometers
distinguishes weekday listings (purple) from weekend listings (magenta). The plot reveals a general negativ
center tend to be less expensive. Both variables are numerical, making a scatterplot the appropriate choic
"""

print(caption_1)
print(caption_2)
print(caption_3)
```

```
Figure 1 - Bar Chart: Average AirBnB Prices by City
This bar chart displays the mean nightly rental price across ten European cities. This display gives us inf
ormation on the relationship between categorical (city on x-axis)
and numerical information (average price on y-axis). London shows the highest average price, followed by Pa
ris
and Amsterdam, while Budapest and Athens offer the most affordable options.


Figure 2 - Histogram: Distribution of AirBnB Prices under 1000
This histogram illustrates the frequency distribution of AirBnB listing prices under 1000 across all ten ci
ties. The distribution is right-skewed,
with most listings concentrated between 50-200 per night and as seen by the dashed line that doesn't line u
p with the highest peak. This visualization is useful for showing
the distribution of a single numerical variable.


Figure 3 - Scatterplot: Price vs. Distance from City Center
This scatterplot examines the relationship between an AirBnB's distance from the city center (in kilometer
s) and its nightly price. Color encoding
distinguishes weekday listings (purple) from weekend listings (magenta). The plot reveals a general negativ
e trend where properties farther from the city
center tend to be less expensive. Both variables are numerical, making a scatterplot the appropriate choic
e.
```