- Jeremy Beard
- Due 20220320
- Week 2
- MSDS650
- Genie Hays

Week 2 Lab - Calculating Probability and Proportion

This week's assignment will give you some practice calculating and interpretting probablity and proportions.

Dataset Name:: flights.csv, airlines.csv, airports.csv (found in the assign_wk2 folder)
You really only need the flights.csv to complete the assignment, however, the other two datasets provide some reference info that you might find interesting.

Since the original flights dataset has **lots** of missing data, I have provided a cleaned up version for you to use (flights_clean.csv). I have also provided the notebook that I used to clean up the dataset (Clean_Flights_Data).

For those of you who wish to try your hand at data cleaning, I have provided a notebook demonstrating data imputation data (Demo_Imputing_Data). You get to decide which version of the dataset you wish to use.

Assignment Requirements

Here are the requirements for this week's assignment:

- Load your choice of dataset (either flights.csv **OR** flights_clean.csv)
  - If you are going to clean the dataset yourself, here are some hints:
    - Warning!! You are going to need some of the rows/columns with missing values, so don't just throw them away while creating your dataframe
    - The column 'ARRIVAL_DELAY' tells you the number of minutes the flight actually arrived verses the scheduled arrival. There are a fair number of missing values for this column, impute (see Demo_Imputing_Data for ideas) this column. Document your approach!
    - Hint: A negative number means the flight arrived early.
    - Hint: What other columns might you use to fill in this missing data.
  - If you are going to start with cleaned_dataset, I encourage you to look at what I did to clean up this data. It will help you going forward

- Provide an analysis of delayed flights based on the airport the flight originated from. Your analysis should answer the following questions. 1) Determine the originaing airport with the largest proportion of flights arriving late to their destination. Do the same for the airport with the smallest proportion. 2) What is the probablity a flight leaving from a given airport will arrive at its destination late?

```
* Hint: Calculate the probablity of late arrival at destination for each originating airport.
```

3) What is the mean and std of late arrival times for both of these airports.

```
* Based on the mean and std information ONLY, which airport seems like a better choice?
```

4) Define a question that would utilize Bernoulli's Equation and preform a calculation to support your question. 5) Provide a summary of all the values that you calculated for 3 airports

```
* Compare the three to each other.
* Which airport would you prefer to fly out of based on your results.
```

Deliverables

Upload your Jupyter Notebook to the corresponding location in WorldClass.

Note:: Make sure you have clearly indicated each assignment requirement within your notebook.

I. Introduction

In this week's assignment, the flights_clean.csv dataset was used (thank you for doing the work cleaning the data!) The main idea of this week's assignment was calculating delayed arrivals from flights coming from different originating airports. I chose to analyze the Denver, Chicago, and Atlanta airports. Basic statistics were calculated on the data and proportions and probabilities were found. Bernoulli's equation was also involved and you will see what question I answered using Bernoulli's equation.

II. Methods, III. Code, and IV. Analysis of Results

First, I load the data. Again, I chose to use the cleaned data as that eliminated a lot of the work I would have to do to clean the data. After loading the data, I then look at the metadata using shape() and info() functions.

In [1]:

```
import pandas as pd
import numpy as np
from scipy import stats

import matplotlib.pylab as plt
%matplotlib inline
```



```
#Load your choice of dataset (either flights.csv OR flights_clean.csv)
```

```
##If you are going to clean the dataset yourself, here are some hints:
###Warning!! You are going to need some of the rows/columns with missing values, so don't just throw them away while creating your c
###The column 'ARRIVAL_DELAY' tells you the number of minutes the flight actually arrived verses the scheduled arrival. There are a
###Hint: A negative number means the flight arrived early.
###Hint: What other columns might you use to fill in this missing data.
##If you are going to start with cleaned_dataset, I encourage you to look at what I did to clean up this data. It will help you goi
data_df = pd.read_csv('assign_wk2/flights_clean.csv')
```

In [3]:

```
data_df.head(10)
```

Out[3]:

	year	month	day	day_of_week	airline	flight_number	origin_airport	destination_airport	scheduled_departure	departure_time	departure_delay	schedu
0	2015	1	1	4	AS	98	ANC	SEA	5	2354.0	-11.0	
1	2015	1	1	4	AA	2336	LAX	PBI	10	2.0	-8.0	
2	2015	1	1	4	US	840	SFO	CLT	20	18.0	-2.0	
3	2015	1	1	4	AA	258	LAX	MIA	20	15.0	-5.0	
4	2015	1	1	4	AS	135	SEA	ANC	25	24.0	-1.0	
5	2015	1	1	4	DL	806	SFO	MSP	25	20.0	-5.0	
6	2015	1	1	4	NK	612	LAS	MSP	25	19.0	-6.0	
7	2015	1	1	4	US	2013	LAX	CLT	30	44.0	14.0	
8	2015	1	1	4	AA	1112	SFO	DFW	30	19.0	-11.0	
9	2015	1	1	4	DL	1173	LAS	ATL	30	33.0	3.0	

In [4]:

```
data_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5245484 entries, 0 to 5245483
Data columns (total 18 columns):
 #   Column               Dtype
---  ------               -----
 0   year                 int64
 1   month                int64
 2   day                  int64
 3   day_of_week          int64
 4   airline              object
 5   flight_number        int64
 6   origin_airport       object
         _airport  object
 8   scheduled_departure  int64
```

Loading [MathJax]/extensions/Safe.js

```
 9   departure_time       float64
 10  departure_delay      float64
 11  scheduled_time       float64
 12  elapsed_time         float64
 13  scheduled_arrival    int64
 14  arrival_time         float64
 15  arrival_delay        float64
 16  diverted             int64
 17  cancelled            int64
dtypes: float64(6), int64(9), object(3)
memory usage: 720.4+ MB
```

In [5]:

```python
data_df.shape
```

Out[5]:

```
(5245484, 18)
```

In [6]:

```python
og_airport_counts = data_df.groupby('origin_airport').size()
og_airport_counts.head(30)
```

Out[6]:

```
origin_airport
ABE      2235
ABI      2232
ABQ     18980
ABR       663
ABY       867
ACK       486
ACT      1539
ACV      1269
ACY      3532
ADK        89
ADQ       437
AEX      3060
AGS      2346
AKN        63
ALB      7341
ALO       582
AMA      4080
ANC     15881
APN       547
ASE      3286
ATL    344279
ATW      2765
AUS     41489
AVL      2684
AVP      1331
AZO      1743
BET       880
```

```
BFL      2595
BGM       259
dtype: int64
```

Question 1

After loading the data and looking at its metadata, I sought to answer the question "what is the originating airport with the largest proportion of flights arriving late?" I first created a 'late' column based on if the flight arrived late or not. This could easily be determined by looking at the arrival_delay column but I wanted to make the information more explicit by using True/False boolean values.

In [7]:

```
#Provide an analysis of delayed flights based on the airport the
#flight originated from. Your analysis should answer the following questions.

#1) Determine the origінaing airport with the largest proportion of flights arriving late to their destination.
##Do the same for the airport with the smallest proportion.
data_df.isnull().sum()
data_df['late'] = data_df.arrival_delay.apply(lambda x: x > 0)
data_df.head(10)
```

Out[7]:

	year	month	day	day_of_week	airline	flight_number	origin_airport	destination_airport	scheduled_departure	departure_time	departure_delay	schedu
0	2015	1	1	4	AS	98	ANC	SEA	5	2354.0	-11.0	
1	2015	1	1	4	AA	2336	LAX	PBI	10	2.0	-8.0	
2	2015	1	1	4	US	840	SFO	CLT	20	18.0	-2.0	
3	2015	1	1	4	AA	258	LAX	MIA	20	15.0	-5.0	
4	2015	1	1	4	AS	135	SEA	ANC	25	24.0	-1.0	
5	2015	1	1	4	DL	806	SFO	MSP	25	20.0	-5.0	
6	2015	1	1	4	NK	612	LAS	MSP	25	19.0	-6.0	
7	2015	1	1	4	US	2013	LAX	CLT	30	44.0	14.0	
8	2015	1	1	4	AA	1112	SFO	DFW	30	19.0	-11.0	
9	2015	1	1	4	DL	1173	LAS	ATL	30	33.0	3.0	

In [8]:

```
lateCounts = data_df[data_df.late == True].groupby('origin_airport').size().sort_values(ascending=False)
lateCounts
```

Out[8]:

```
origin_airport
ATL      115034
ORD      113415
```

Loading [MathJax]/extensions/Safe.js

```
DFW     93195
DEN     80870
LAX     79754
        ...
VEL        23
HYA        21
DLG        17
ITH        14
CNY         8
Length: 322, dtype: int64
```

In [9]:

```
notLateCounts = data_df[data_df.late == False].groupby('origin_airport').size().sort_values(ascending=False)
notLateCounts
```

Out[9]:

```
origin_airport
ATL    229245
ORD    163923
DFW    140102
DEN    113062
LAX    112755
        ...
PPG        40
ADK        32
AKN        31
GST        25
ITH        16
Length: 322, dtype: int64
```

I calculated the proportion of late arrivals based on originating airport by counting the number of late arrivals based on the origin_airport column. I then added a column that divides the number of late counts for the airport by the total number of flight counts for that airport.

In [10]:

```
og_airport_lateCounts = data_df.groupby(['origin_airport', 'late']).size().unstack().reset_index()
cols = ['origin_airport', 'not_late', 'late']
og_airport_lateCounts.columns = cols
og_airport_lateCounts['total'] = og_airport_lateCounts.not_late + og_airport_lateCounts.late
og_airport_lateCounts['late_prop'] = og_airport_lateCounts.late/og_airport_lateCounts.total
og_airport_lateCounts.head(30)
```

Out[10]:

	origin_airport	not_late	late	total	late_prop
0	ABE	1409	826	2235	0.369575
1	ABI	1546	686	2232	0.307348
2	ABQ	12008	6972	18980	0.367334
3	ABR	417	246	663	0.371041
4	ABY	558	309	867	0.356401

	origin_airport	not_late	late	total	late_prop
5	ACK	324	162	486	0.333333
6	ACT	1073	466	1539	0.302794
7	ACV	820	449	1269	0.353822
8	ACY	2099	1433	3532	0.405719
9	ADK	32	57	89	0.640449
10	ADQ	277	160	437	0.366133
11	AEX	1881	1179	3060	0.385294
12	AGS	1342	1004	2346	0.427962
13	AKN	31	32	63	0.507937
14	ALB	5335	2006	7341	0.273260
15	ALO	410	172	582	0.295533
16	AMA	2565	1515	4080	0.371324
17	ANC	10667	5214	15881	0.328317
18	APN	405	142	547	0.259598
19	ASE	1822	1464	3286	0.445526
20	ATL	229245	115034	344279	0.334130
21	ATW	1688	1077	2765	0.389512
22	AUS	26797	14692	41489	0.354118
23	AVL	1750	934	2684	0.347988
24	AVP	868	463	1331	0.347859
25	AZO	1245	498	1743	0.285714
26	BDL	12739	5698	18437	0.309052
27	BET	552	328	880	0.372727
28	BFL	1744	851	2595	0.327938
29	BGM	189	70	259	0.270270

After calculating the proportions, I sorted the proportion values by increasing and decreasing values. As you can see, the highest proportion of late arrivals was found in GST airport, in Gustavus, Alaska. The lowest proportion of late arrivals was found at the CNY airport, the Canyonlands Regional Airport.

In [11]:

```python
og_airport_lateCounts = og_airport_lateCounts.sort_values('late_prop', ascending=False)
og_airport_lateCounts.head(30)
```

Out[11]:

	origin_airport	not_late	late	total	late_prop
133	GST	25	51	76	0.671053
9	ADK	32	57	89	0.640449
246	PPG	40	67	107	0.626168
154	ILG	42	53	95	0.557895
13	AKN	31	32	63	0.507937
258	RHI	479	460	939	0.489883
226	OME	325	311	636	0.488994
293	STC	40	38	78	0.487179
71	COD	340	313	653	0.479326
90	DLH	889	778	1667	0.466707
161	ITH	16	14	30	0.466667
232	OTZ	342	299	641	0.466459
234	PBG	149	130	279	0.465950
81	DAL	32223	26626	58849	0.452446
62	CIU	328	269	597	0.450586
40	BPT	503	407	910	0.447253
19	ASE	1822	1464	3286	0.445526
76	CRW	1292	1034	2326	0.444540
222	OAK	23324	18419	41743	0.441248
47	BTR	3926	3077	7003	0.439383
123	GGG	347	268	615	0.435772
231	OTH	150	115	265	0.433962
87	DHN	705	538	1243	0.432824
143	HOU	29285	21983	51268	0.428786
[illegible]	[illegible]	1342	1004	2346	0.427962

	origin_airport	not_late	late	total	late_prop
313	UST	83	61	144	0.423611
144	HPN	4167	2997	7164	0.418342
66	CLT	57761	41482	99243	0.417984
85	DEN	113062	80870	193932	0.417002
319	XNA	5259	3728	8987	0.414821

In [12]:

```
og_airport_lateCounts = og_airport_lateCounts.sort_values('late_prop', ascending=True)
og_airport_lateCounts.head(20)
```

Out[12]:

	origin_airport	not_late	late	total	late_prop
70	CNY	197	8	205	0.039024
314	VEL	177	23	200	0.115000
46	BTM	549	99	648	0.152778
189	LWS	491	97	588	0.164966
88	DIK	730	191	921	0.207383
98	EKO	405	112	517	0.216634
89	DLG	60	17	77	0.220779
32	BIL	2220	633	2853	0.221872
119	GCC	766	220	986	0.223124
102	ESC	430	126	556	0.226619
80	DAB	1152	338	1490	0.226846
74	CPR	1348	397	1745	0.227507
260	RKS	519	153	672	0.227679
96	ECP	3184	965	4149	0.232586
140	HLN	1089	337	1426	0.236325
173	LAR	425	132	557	0.236984
153	IDA	1688	540	2228	0.242370
302	TOL	677	220	897	0.245262

	origin_airport	not_late	late	total	late_prop
255	RDD	532	174	706	0.246459
320	YAK	495	162	657	0.246575

Question 2

After answering the questions on proportion, I then sought to answer the question "What is the probablity a flight leaving from a given airport will arrive at its destination late?" I mainly followed the SampleAssignment provided by Professor Hayes. Since a probability is based on the total number of outcomes, these probabilities were calculated based on the total number of flights available from the dataset.

In [13]:

```
#2) What is the probablity a flight leaving from a given airport will arrive at its destination late?
##  * Hint: Calculate the probablity of late arrival at destination for each originating airport.

lateFlights = data_df[data_df.late == True].groupby('origin_airport').size()
lateFlights
```

Out[13]:

```
origin_airport
ABE      826
ABI      686
ABQ     6972
ABR      246
ABY      309
        ...
WRG      236
WYS       58
XNA     3728
YAK      162
YUM      627
Length: 322, dtype: int64
```

In [14]:

```
# I will now calculate the probabilities of late arrivals for all origin_airports
# source: SampleAssignment_Week1_Hays
data_df[(data_df.late == True) & (data_df.origin_airport == 'DEN')]
```

Out[14]:

	year	month	day	day_of_week	airline	flight_number	origin_airport	destination_airport	scheduled_departure	departure_time	departure_delay
29	2015	1	1	4	AA	2392	DEN	MIA	120	141.0	21.0
82	2015	1	1	4	AA	328	DEN	DFW	530	623.0	53.0
123	2015	1	1	4	OO	2599	DEN	LAX	545	658.0	73.0
568	2015	1	1	4	F9	1246	DEN	DFW	630	634.0	4.0
672	2015	1	1	4	F9	110	DEN	MSP	645	711.0	26.0

Loading [MathJax]/extensions/Safe.js

	year	month	day	day_of_week	airline	flight_number	origin_airport	destination_airport	scheduled_departure	departure_time	departure_delay
...	...	...	...	...	...	...	...	...	...	...	...
5244861	2015	12	31	4	F9	242	DEN	IAH	2040	2054.0	14.0
5244929	2015	12	31	4	F9	332	DEN	MSP	2055	2125.0	30.0
5244939	2015	12	31	4	WN	5215	DEN	TUS	2055	2109.0	14.0
5245141	2015	12	31	4	F9	761	DEN	LAS	2148	2141.0	-7.0
5245399	2015	12	31	4	B6	994	DEN	BOS	2318	2349.0	31.0

80870 rows × 19 columns

In [15]:

```
total_flights = len(data_df)
total_flights
```

Out[15]:

```
5245484
```

In [16]:

```
prob_delay_airport = lateFlights.apply(lambda x: x/total_flights)
prob_delay_airport = prob_delay_airport.sort_values(ascending=False)
prob_delay_airport
```

Out[16]:

```
origin_airport
ATL    0.021930
ORD    0.021621
DFW    0.017767
DEN    0.015417
LAX    0.015204
         ...
VEL    0.000004
HYA    0.000004
DLG    0.000003
ITH    0.000003
CNY    0.000002
Length: 322, dtype: float64
```

In [17]:

```
# a bit nicer output for all the origin_airports (still following the example, thank you!)
for i in prob_delay_airport.items():
    p_delay = '%.6f'%(i[1]*100)
    print(f'A flight from {i[0]} has a {p_delay}% probability of having a delayed arrival.')
```

Loading [MathJax]/extensions/Safe.js

```
L has a 2.193010% probability of having a delayed arrival.
A flight from ORD has a 2.162107% probability of having a delayed arrival.
```

```
A flight from DFW has a 1.776671% probability of having a delayed arrival.
A flight from DEN has a 1.541707% probability of having a delayed arrival.
A flight from LAX has a 1.520432% probability of having a delayed arrival.
A flight from IAH has a 1.142011% probability of having a delayed arrival.
A flight from PHX has a 1.112462% probability of having a delayed arrival.
A flight from SFO has a 1.072179% probability of having a delayed arrival.
A flight from LAS has a 1.009497% probability of having a delayed arrival.
A flight from SEA has a 0.831096% probability of having a delayed arrival.
A flight from CLT has a 0.790814% probability of having a delayed arrival.
A flight from MCO has a 0.788869% probability of having a delayed arrival.
A flight from BOS has a 0.740065% probability of having a delayed arrival.
A flight from MSP has a 0.716731% probability of having a delayed arrival.
A flight from DTW has a 0.692539% probability of having a delayed arrival.
A flight from LGA has a 0.691604% probability of having a delayed arrival.
A flight from EWR has a 0.687296% probability of having a delayed arrival.
A flight from BWI has a 0.667546% probability of having a delayed arrival.
A flight from JFK has a 0.639083% probability of having a delayed arrival.
A flight from MDW has a 0.608180% probability of having a delayed arrival.
A flight from SLC has a 0.594130% probability of having a delayed arrival.
A flight from MIA has a 0.529522% probability of having a delayed arrival.
A flight from DAL has a 0.507599% probability of having a delayed arrival.
A flight from SAN has a 0.478354% probability of having a delayed arrival.
A flight from FLL has a 0.473055% probability of having a delayed arrival.
A flight from DCA has a 0.464361% probability of having a delayed arrival.
A flight from PHL has a 0.463237% probability of having a delayed arrival.
A flight from TPA has a 0.421601% probability of having a delayed arrival.
A flight from HOU has a 0.419084% probability of having a delayed arrival.
A flight from OAK has a 0.351140% probability of having a delayed arrival.
A flight from STL has a 0.320676% probability of having a delayed arrival.
A flight from BNA has a 0.319551% probability of having a delayed arrival.
A flight from HNL has a 0.295073% probability of having a delayed arrival.
A flight from PDX has a 0.288248% probability of having a delayed arrival.
A flight from SJC has a 0.287218% probability of having a delayed arrival.
A flight from AUS has a 0.280089% probability of having a delayed arrival.
A flight from SMF has a 0.279459% probability of having a delayed arrival.
A flight from MCI has a 0.256964% probability of having a delayed arrival.
A flight from MSY has a 0.243886% probability of having a delayed arrival.
A flight from SNA has a 0.238205% probability of having a delayed arrival.
A flight from IAD has a 0.231494% probability of having a delayed arrival.
A flight from CLE has a 0.209361% probability of having a delayed arrival.
A flight from RDU has a 0.207664% probability of having a delayed arrival.
A flight from MKE has a 0.182443% probability of having a delayed arrival.
A flight from SAT has a 0.181108% probability of having a delayed arrival.
A flight from RSW has a 0.178268% probability of having a delayed arrival.
A flight from PBI has a 0.168812% probability of having a delayed arrival.
A flight from SJU has a 0.152951% probability of having a delayed arrival.
A flight from IND has a 0.151997% probability of having a delayed arrival.
A flight from OGG has a 0.150282% probability of having a delayed arrival.
H has a 0.147193% probability of having a delayed arrival.
A flight from CVG has a 0.133353% probability of having a delayed arrival.
```

```
A flight from ABQ has a 0.132914% probability of having a delayed arrival.
A flight from PIT has a 0.129616% probability of having a delayed arrival.
A flight from ONT has a 0.117816% probability of having a delayed arrival.
A flight from BUR has a 0.115032% probability of having a delayed arrival.
A flight from JAX has a 0.110438% probability of having a delayed arrival.
A flight from BDL has a 0.108627% probability of having a delayed arrival.
A flight from OMA has a 0.106701% probability of having a delayed arrival.
A flight from OKC has a 0.103804% probability of having a delayed arrival.
A flight from ANC has a 0.099400% probability of having a delayed arrival.
A flight from RIC has a 0.097169% probability of having a delayed arrival.
A flight from TUS has a 0.093185% probability of having a delayed arrival.
A flight from MEM has a 0.092003% probability of having a delayed arrival.
A flight from RNO has a 0.091183% probability of having a delayed arrival.
A flight from BUF has a 0.091145% probability of having a delayed arrival.
A flight from TUL has a 0.088934% probability of having a delayed arrival.
A flight from ELP has a 0.082204% probability of having a delayed arrival.
A flight from BHM has a 0.080660% probability of having a delayed arrival.
A flight from BOI has a 0.079116% probability of having a delayed arrival.
A flight from XNA has a 0.071071% probability of having a delayed arrival.
A flight from KOA has a 0.071052% probability of having a delayed arrival.
A flight from CHS has a 0.069641% probability of having a delayed arrival.
A flight from LIT has a 0.069241% probability of having a delayed arrival.
A flight from LIH has a 0.064684% probability of having a delayed arrival.
A flight from PSP has a 0.063064% probability of having a delayed arrival.
A flight from PVD has a 0.062892% probability of having a delayed arrival.
A flight from SDF has a 0.062454% probability of having a delayed arrival.
A flight from GRR has a 0.061081% probability of having a delayed arrival.
A flight from GEG has a 0.060338% probability of having a delayed arrival.
A flight from BTR has a 0.058660% probability of having a delayed arrival.
A flight from HPN has a 0.057135% probability of having a delayed arrival.
A flight from LGB has a 0.056735% probability of having a delayed arrival.
A flight from ORF has a 0.056163% probability of having a delayed arrival.
A flight from DAY has a 0.055553% probability of having a delayed arrival.
A flight from DSM has a 0.055305% probability of having a delayed arrival.
A flight from ICT has a 0.054161% probability of having a delayed arrival.
A flight from FAT has a 0.053570% probability of having a delayed arrival.
A flight from MSN has a 0.052617% probability of having a delayed arrival.
A flight from TYS has a 0.050825% probability of having a delayed arrival.
A flight from SAV has a 0.049624% probability of having a delayed arrival.
A flight from MAF has a 0.049490% probability of having a delayed arrival.
A flight from CID has a 0.048613% probability of having a delayed arrival.
A flight from COS has a 0.047145% probability of having a delayed arrival.
A flight from SGF has a 0.046554% probability of having a delayed arrival.
A flight from SHV has a 0.044667% probability of having a delayed arrival.
A flight from JAN has a 0.044667% probability of having a delayed arrival.
A flight from GSO has a 0.043180% probability of having a delayed arrival.
A flight from ROC has a 0.040549% probability of having a delayed arrival.
A flight from GSP has a 0.040301% probability of having a delayed arrival.
Loading [MathJax]/extensions/Safe.js D has a 0.040244% probability of having a delayed arrival.
A flight from PNS has a 0.038547% probability of having a delayed arrival.
```

```
A flight from ALB has a 0.038242% probability of having a delayed arrival.
A flight from MOB has a 0.038185% probability of having a delayed arrival.
A flight from FAR has a 0.037937% probability of having a delayed arrival.
A flight from SBA has a 0.037670% probability of having a delayed arrival.
A flight from LFT has a 0.037232% probability of having a delayed arrival.
A flight from FWA has a 0.036260% probability of having a delayed arrival.
A flight from ITO has a 0.035612% probability of having a delayed arrival.
A flight from CAE has a 0.035593% probability of having a delayed arrival.
A flight from MYR has a 0.035002% probability of having a delayed arrival.
A flight from LEX has a 0.034201% probability of having a delayed arrival.
A flight from MHT has a 0.033801% probability of having a delayed arrival.
A flight from CRP has a 0.032104% probability of having a delayed arrival.
A flight from GRB has a 0.032008% probability of having a delayed arrival.
A flight from CAK has a 0.031684% probability of having a delayed arrival.
A flight from SBN has a 0.031684% probability of having a delayed arrival.
A flight from PIA has a 0.031627% probability of having a delayed arrival.
A flight from JNU has a 0.030846% probability of having a delayed arrival.
A flight from SYR has a 0.030388% probability of having a delayed arrival.
A flight from LBB has a 0.029568% probability of having a delayed arrival.
A flight from VPS has a 0.029530% probability of having a delayed arrival.
A flight from CHA has a 0.029034% probability of having a delayed arrival.
A flight from AMA has a 0.028882% probability of having a delayed arrival.
A flight from ASE has a 0.027910% probability of having a delayed arrival.
A flight from PWM has a 0.027910% probability of having a delayed arrival.
A flight from ACY has a 0.027319% probability of having a delayed arrival.
A flight from HSV has a 0.026918% probability of having a delayed arrival.
A flight from MLI has a 0.026575% probability of having a delayed arrival.
A flight from ISP has a 0.026537% probability of having a delayed arrival.
A flight from STT has a 0.026003% probability of having a delayed arrival.
A flight from EVV has a 0.025736% probability of having a delayed arrival.
A flight from GRK has a 0.025145% probability of having a delayed arrival.
A flight from MGM has a 0.024440% probability of having a delayed arrival.
A flight from FNT has a 0.023392% probability of having a delayed arrival.
A flight from BIS has a 0.022839% probability of having a delayed arrival.
A flight from AEX has a 0.022476% probability of having a delayed arrival.
A flight from MLU has a 0.021733% probability of having a delayed arrival.
A flight from TTN has a 0.021485% probability of having a delayed arrival.
A flight from RAP has a 0.021295% probability of having a delayed arrival.
A flight from EUG has a 0.021199% probability of having a delayed arrival.
A flight from BZN has a 0.021123% probability of having a delayed arrival.
A flight from SBP has a 0.020723% probability of having a delayed arrival.
A flight from ATW has a 0.020532% probability of having a delayed arrival.
A flight from GNV has a 0.020189% probability of having a delayed arrival.
A flight from BTV has a 0.019941% probability of having a delayed arrival.
A flight from GPT has a 0.019827% probability of having a delayed arrival.
A flight from MRY has a 0.019750% probability of having a delayed arrival.
A flight from CRW has a 0.019712% probability of having a delayed arrival.
A flight from AGS has a 0.019140% probability of having a delayed arrival.
[illegible]C has a 0.019007% probability of having a delayed arrival.
A flight from SRQ has a 0.018873% probability of having a delayed arrival.
```

```
A flight from ECP has a 0.018397% probability of having a delayed arrival.
A flight from MFE has a 0.018282% probability of having a delayed arrival.
A flight from TLH has a 0.018263% probability of having a delayed arrival.
A flight from AVL has a 0.017806% probability of having a delayed arrival.
A flight from GJT has a 0.017787% probability of having a delayed arrival.
A flight from BMI has a 0.017444% probability of having a delayed arrival.
A flight from LNK has a 0.016853% probability of having a delayed arrival.
A flight from KTN has a 0.016776% probability of having a delayed arrival.
A flight from MDT has a 0.016605% probability of having a delayed arrival.
A flight from BFL has a 0.016223% probability of having a delayed arrival.
A flight from HRL has a 0.016147% probability of having a delayed arrival.
A flight from ABE has a 0.015747% probability of having a delayed arrival.
A flight from TVC has a 0.015728% probability of having a delayed arrival.
A flight from TYR has a 0.015575% probability of having a delayed arrival.
A flight from ROA has a 0.015042% probability of having a delayed arrival.
A flight from DLH has a 0.014832% probability of having a delayed arrival.
A flight from DRO has a 0.014565% probability of having a delayed arrival.
A flight from ELM has a 0.014565% probability of having a delayed arrival.
A flight from MFR has a 0.014470% probability of having a delayed arrival.
A flight from ISN has a 0.014336% probability of having a delayed arrival.
A flight from CHO has a 0.014298% probability of having a delayed arrival.
A flight from PSC has a 0.013650% probability of having a delayed arrival.
A flight from FAI has a 0.013154% probability of having a delayed arrival.
A flight from ABI has a 0.013078% probability of having a delayed arrival.
A flight from FSM has a 0.013002% probability of having a delayed arrival.
A flight from LRD has a 0.012697% probability of having a delayed arrival.
A flight from FLG has a 0.012678% probability of having a delayed arrival.
A flight from RST has a 0.012563% probability of having a delayed arrival.
A flight from CLL has a 0.012449% probability of having a delayed arrival.
A flight from LAN has a 0.012392% probability of having a delayed arrival.
A flight from BRO has a 0.012258% probability of having a delayed arrival.
A flight from BIL has a 0.012068% probability of having a delayed arrival.
A flight from TRI has a 0.012029% probability of having a delayed arrival.
A flight from YUM has a 0.011953% probability of having a delayed arrival.
A flight from CMI has a 0.011705% probability of having a delayed arrival.
A flight from MSO has a 0.011095% probability of having a delayed arrival.
A flight from MOT has a 0.011057% probability of having a delayed arrival.
A flight from FAY has a 0.010866% probability of having a delayed arrival.
A flight from SAF has a 0.010790% probability of having a delayed arrival.
A flight from ILM has a 0.010752% probability of having a delayed arrival.
A flight from RDM has a 0.010752% probability of having a delayed arrival.
A flight from SGU has a 0.010428% probability of having a delayed arrival.
A flight from IDA has a 0.010295% probability of having a delayed arrival.
A flight from DHN has a 0.010256% probability of having a delayed arrival.
A flight from FCA has a 0.010028% probability of having a delayed arrival.
A flight from MBS has a 0.009970% probability of having a delayed arrival.
A flight from GTF has a 0.009761% probability of having a delayed arrival.
A flight from EYW has a 0.009704% probability of having a delayed arrival.
K has a 0.009704% probability of having a delayed arrival.
A flight from SJT has a 0.009589% probability of having a delayed arrival.
```

```
A flight from SPI has a 0.009589% probability of having a delayed arrival.
A flight from AZO has a 0.009494% probability of having a delayed arrival.
A flight from LCH has a 0.009399% probability of having a delayed arrival.
A flight from PHF has a 0.009227% probability of having a delayed arrival.
A flight from LBE has a 0.009017% probability of having a delayed arrival.
A flight from EGE has a 0.008903% probability of having a delayed arrival.
A flight from COU has a 0.008884% probability of having a delayed arrival.
A flight from ACT has a 0.008884% probability of having a delayed arrival.
A flight from SIT has a 0.008827% probability of having a delayed arrival.
A flight from AVP has a 0.008827% probability of having a delayed arrival.
A flight from RHI has a 0.008769% probability of having a delayed arrival.
A flight from BQN has a 0.008655% probability of having a delayed arrival.
A flight from ACV has a 0.008560% probability of having a delayed arrival.
A flight from LAW has a 0.008178% probability of having a delayed arrival.
A flight from CWA has a 0.008007% probability of having a delayed arrival.
A flight from LSE has a 0.007988% probability of having a delayed arrival.
A flight from OAJ has a 0.007931% probability of having a delayed arrival.
A flight from SPS has a 0.007854% probability of having a delayed arrival.
A flight from BPT has a 0.007759% probability of having a delayed arrival.
A flight from CPR has a 0.007568% probability of having a delayed arrival.
A flight from CSG has a 0.007282% probability of having a delayed arrival.
A flight from STX has a 0.006882% probability of having a delayed arrival.
A flight from VLD has a 0.006844% probability of having a delayed arrival.
A flight from GTR has a 0.006691% probability of having a delayed arrival.
A flight from MLB has a 0.006596% probability of having a delayed arrival.
A flight from DAB has a 0.006444% probability of having a delayed arrival.
A flight from HLN has a 0.006425% probability of having a delayed arrival.
A flight from ROW has a 0.006367% probability of having a delayed arrival.
A flight from BET has a 0.006253% probability of having a delayed arrival.
A flight from TXK has a 0.006234% probability of having a delayed arrival.
A flight from MEI has a 0.006177% probability of having a delayed arrival.
A flight from COD has a 0.005967% probability of having a delayed arrival.
A flight from OME has a 0.005929% probability of having a delayed arrival.
A flight from ABY has a 0.005891% probability of having a delayed arrival.
A flight from MTJ has a 0.005872% probability of having a delayed arrival.
A flight from SCC has a 0.005738% probability of having a delayed arrival.
A flight from OTZ has a 0.005700% probability of having a delayed arrival.
A flight from BRW has a 0.005681% probability of having a delayed arrival.
A flight from BQK has a 0.005586% probability of having a delayed arrival.
A flight from ERI has a 0.005414% probability of having a delayed arrival.
A flight from SCE has a 0.005338% probability of having a delayed arrival.
A flight from GFK has a 0.005185% probability of having a delayed arrival.
A flight from CIU has a 0.005128% probability of having a delayed arrival.
A flight from GGG has a 0.005109% probability of having a delayed arrival.
A flight from JMS has a 0.004976% probability of having a delayed arrival.
A flight from JLN has a 0.004957% probability of having a delayed arrival.
A flight from SWF has a 0.004957% probability of having a delayed arrival.
A flight from HDN has a 0.004919% probability of having a delayed arrival.
Loading [MathJax]/extensions/Safe.js
A flight from [illegible]G has a 0.004880% probability of having a delayed arrival.
A flight from HIB has a 0.004861% probability of having a delayed arrival.
```

```
A flight from SUN has a 0.004842% probability of having a delayed arrival.
A flight from ABR has a 0.004690% probability of having a delayed arrival.
A flight from PLN has a 0.004537% probability of having a delayed arrival.
A flight from WRG has a 0.004499% probability of having a delayed arrival.
A flight from DBQ has a 0.004499% probability of having a delayed arrival.
A flight from MKG has a 0.004480% probability of having a delayed arrival.
A flight from BLI has a 0.004404% probability of having a delayed arrival.
A flight from GCK has a 0.004308% probability of having a delayed arrival.
A flight from CLD has a 0.004289% probability of having a delayed arrival.
A flight from GCC has a 0.004194% probability of having a delayed arrival.
A flight from TOL has a 0.004194% probability of having a delayed arrival.
A flight from EAU has a 0.004099% probability of having a delayed arrival.
A flight from PIH has a 0.004080% probability of having a delayed arrival.
A flight from TWF has a 0.004061% probability of having a delayed arrival.
A flight from EWN has a 0.003965% probability of having a delayed arrival.
A flight from PIB has a 0.003965% probability of having a delayed arrival.
A flight from CDV has a 0.003908% probability of having a delayed arrival.
A flight from CMX has a 0.003870% probability of having a delayed arrival.
A flight from SMX has a 0.003870% probability of having a delayed arrival.
A flight from BRD has a 0.003832% probability of having a delayed arrival.
A flight from PAH has a 0.003717% probability of having a delayed arrival.
A flight from ORH has a 0.003679% probability of having a delayed arrival.
A flight from HYS has a 0.003679% probability of having a delayed arrival.
A flight from DIK has a 0.003641% probability of having a delayed arrival.
A flight from INL has a 0.003622% probability of having a delayed arrival.
A flight from GUC has a 0.003603% probability of having a delayed arrival.
A flight from SUX has a 0.003603% probability of having a delayed arrival.
A flight from PSE has a 0.003470% probability of having a delayed arrival.
A flight from CDC has a 0.003432% probability of having a delayed arrival.
A flight from RDD has a 0.003317% probability of having a delayed arrival.
A flight from ALO has a 0.003279% probability of having a delayed arrival.
A flight from BJI has a 0.003279% probability of having a delayed arrival.
A flight from DVL has a 0.003222% probability of having a delayed arrival.
A flight from GRI has a 0.003126% probability of having a delayed arrival.
A flight from IMT has a 0.003107% probability of having a delayed arrival.
A flight from ACK has a 0.003088% probability of having a delayed arrival.
A flight from YAK has a 0.003088% probability of having a delayed arrival.
A flight from ADQ has a 0.003050% probability of having a delayed arrival.
A flight from RKS has a 0.002917% probability of having a delayed arrival.
A flight from HOB has a 0.002802% probability of having a delayed arrival.
A flight from APN has a 0.002707% probability of having a delayed arrival.
A flight from GUM has a 0.002593% probability of having a delayed arrival.
A flight from LAR has a 0.002516% probability of having a delayed arrival.
A flight from PBG has a 0.002478% probability of having a delayed arrival.
A flight from IAG has a 0.002421% probability of having a delayed arrival.
A flight from ESC has a 0.002402% probability of having a delayed arrival.
A flight from BGR has a 0.002288% probability of having a delayed arrival.
A flight from OTH has a 0.002192% probability of having a delayed arrival.
[illegible]O has a 0.002135% probability of having a delayed arrival.
A flight from BTM has a 0.001887% probability of having a delayed arrival.
```

```
A flight from LWS has a 0.001849% probability of having a delayed arrival.
A flight from MQT has a 0.001735% probability of having a delayed arrival.
A flight from PUB has a 0.001659% probability of having a delayed arrival.
A flight from MVY has a 0.001334% probability of having a delayed arrival.
A flight from BGM has a 0.001334% probability of having a delayed arrival.
A flight from PPG has a 0.001277% probability of having a delayed arrival.
A flight from UST has a 0.001163% probability of having a delayed arrival.
A flight from CEC has a 0.001125% probability of having a delayed arrival.
A flight from WYS has a 0.001106% probability of having a delayed arrival.
A flight from ADK has a 0.001087% probability of having a delayed arrival.
A flight from MMH has a 0.001087% probability of having a delayed arrival.
A flight from ILG has a 0.001010% probability of having a delayed arrival.
A flight from GST has a 0.000972% probability of having a delayed arrival.
A flight from STC has a 0.000724% probability of having a delayed arrival.
A flight from AKN has a 0.000610% probability of having a delayed arrival.
A flight from VEL has a 0.000438% probability of having a delayed arrival.
A flight from HYA has a 0.000400% probability of having a delayed arrival.
A flight from DLG has a 0.000324% probability of having a delayed arrival.
A flight from ITH has a 0.000267% probability of having a delayed arrival.
A flight from CNY has a 0.000153% probability of having a delayed arrival.
```

Question 3

After computing the probabilities, I moved onto question 3: "What is the mean and std of late arrival times for both of these airports?" Since question 5 deals with comparing 3 airports, I decided to find the mean and std of late arrival times for 3 airports: Denver, Chicago, and Atlanta. To compute these basic statistics, I filtered the data by only delayed arrival times, and then used the describe() function.

In [18]:

```
#3) What is the mean and std of late arrival times for both of these airports?
##  * Based on the mean and std information ONLY, which airport seems like a better choice?

#For question #3, I will choose 3 airports: DEN, ORD, and ATL, Denver, Chicago, and Atlanta.
#I am choosing 3 airports so question #5 will be easier to answer
#I notice that the question only specifies mean/std of LATE arrival times.
#Therefore, I will only consider the data from DEN which are late arrival times.
delay_stat = data_df[(data_df.late == True)].groupby('origin_airport').arrival_delay.describe()
delay_stat
```

Out[18]:

	count	mean	std	min	25%	50%	75%	max
origin_airport								
ABE	826.0	36.483051	63.478274	1.0	5.00	14.0	35.0	612.0
ABI	686.0	39.295918	63.122970	1.0	6.00	15.0	46.0	583.0
ABQ	6972.0	33.444923	82.563811	1.0	6.00	14.0	35.0	2194.0
[illegible]	[illegible]	42.211382	112.170443	1.0	6.00	13.5	29.0	916.0

Loading [MathJax]/extensions/Safe.js

	count	mean	std	min	25%	50%	75%	max
origin_airport								
ABY	309.0	37.530744	58.046303	1.0	5.00	15.0	46.0	454.0
...	...	...	...	...	...	...	...	...
WRG	236.0	31.881356	45.870710	1.0	5.00	14.0	35.0	259.0
WYS	58.0	18.034483	37.468099	1.0	2.25	5.0	14.5	194.0
XNA	3728.0	43.436427	77.550802	1.0	6.00	17.0	49.0	2100.0
YAK	162.0	25.783951	37.628904	1.0	5.25	13.0	31.0	284.0
YUM	627.0	24.473684	49.477628	1.0	4.00	8.0	19.0	458.0

322 rows × 8 columns

In [19]:

```
# Now will filter delay_stat to find mean/std for DEN, ORD, and ATL
delay_stat.loc['DEN']
```

Out[19]:

```
count    80870.000000
mean        34.795994
std         76.862392
min          1.000000
25%          6.000000
50%         16.000000
75%         40.000000
max       2308.000000
Name: DEN, dtype: float64
```

In [20]:

```
delay_stat.loc['ORD']
```

Out[20]:

```
count    113413.000000
mean         40.924092
std          74.904126
min           1.000000
25%           8.000000
50%          20.000000
75%          50.000000
max        2287.000000
Name: ORD, dtype: float64
```

In [21]:

```
delay_stat.loc['ATL']
```

Loading [MathJax]/extensions/Safe.js

Out[21]:

```
.000000
```

```
mean          32.920658
std           76.114586
min            1.000000
25%            5.000000
50%           14.000000
75%           36.000000
max         2276.000000
Name: ATL, dtype: float64
```

In [22]:

```
#Based on these 3 airports data, the Atlanta airport (ATL) has
# a lower mean delay than either the Chicago airport (ORD) or the Denver airport (DEN)
#Therefore, I would rather have a flight originating from ATL than ORD or DEN.
#All 3 of these airports have similar standard deviations so not much consideration
#was given to the standard deviations of the 3 airports.
```

Aside: Histograms

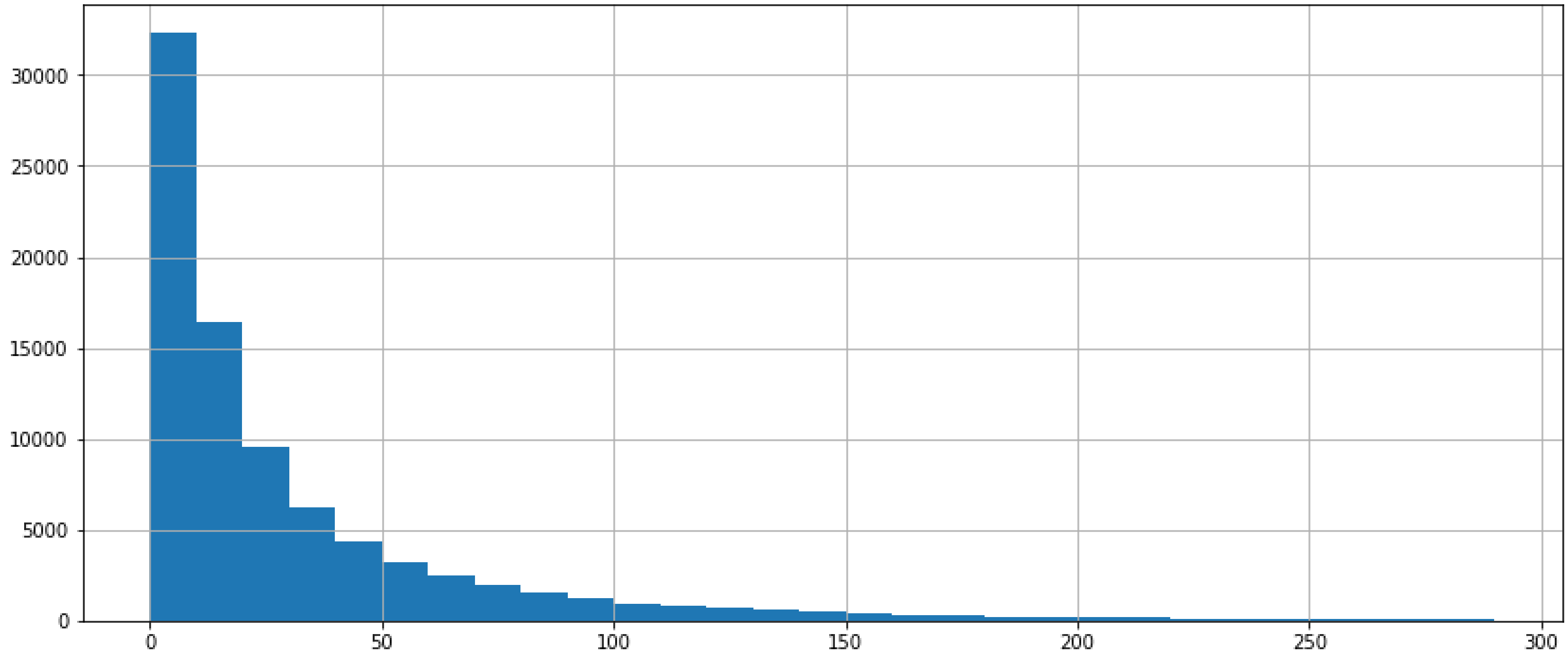
After comparing the means and standard deviations of the 3 airports, I saw that the Atlanta airport had the lowest delay for arrival time. That seems to be the preferable airport when only considering the means and standard deviations. The example I was following went on to create histograms of how much the flight arrived delayed. I wanted to also visualize the data a bit so I created histograms of each of the 3 airports in question.

In [23]:

```
bin_values = np.arange(start=0, stop=300, step=10)
den_delays = data_df[(data_df.origin_airport == 'DEN')]
den_delays.arrival_delay.hist(bins=bin_values, figsize=[14,6])
```

Out[23]:

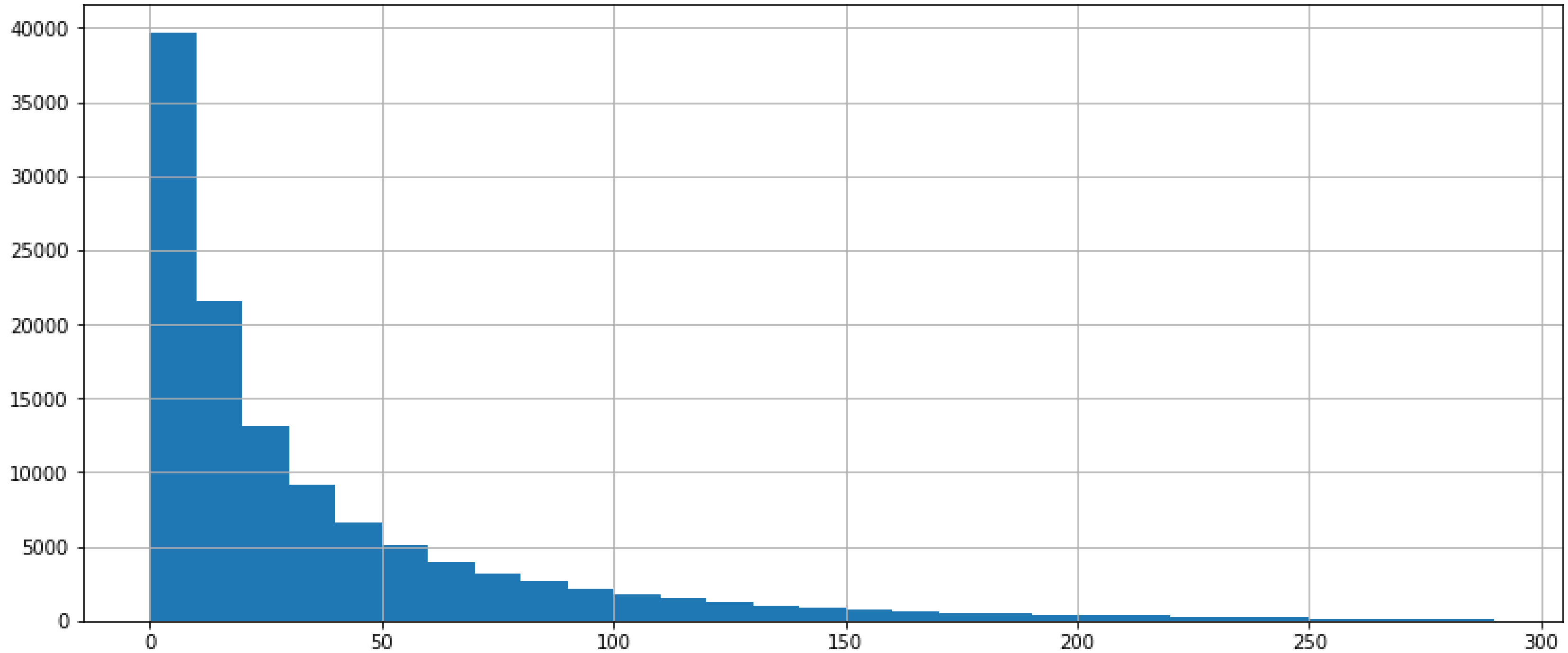
```
<AxesSubplot:>
```



In [24]:

```
ord_delays = data_df[(data_df.origin_airport == 'ORD')]
ord_delays.arrival_delay.hist(bins=bin_values, figsize=[14,6])
```

Out[24]: <AxesSubplot:>

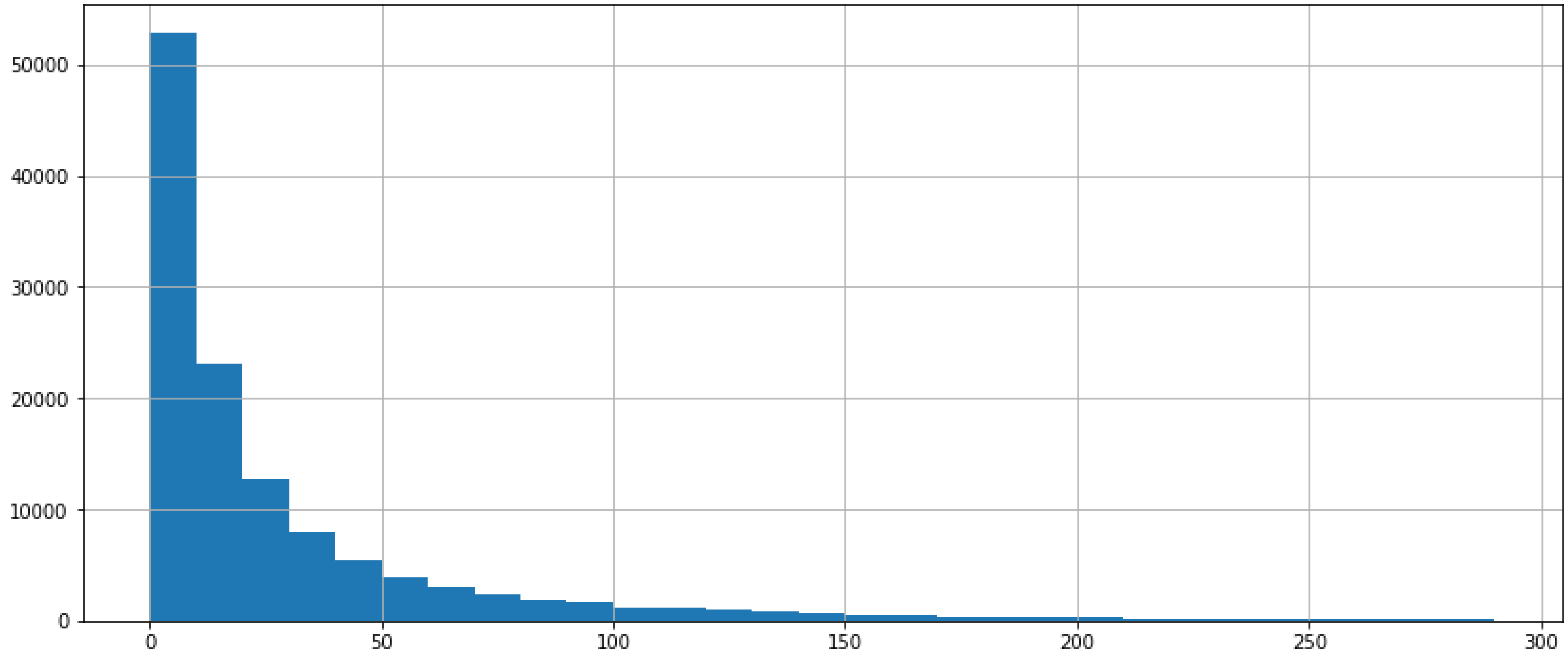


In [25]:

```
atl_delays = data_df[(data_df.origin_airport == 'ATL')]
atl_delays.arrival_delay.hist(bins=bin_values, figsize=[14,6])
```

Out[25]: <AxesSubplot:>

Loading [MathJax]/extensions/Safe.js



Question 4

Question 4 dealt with utilizing Bernoulli's equation. I wanted to see what the probabilities were of there being either 3% or 0.5% of flights per day delayed at the airport in question. I computed these 2 probabilities for each of the 3 airports using Bernoulli's equation.

In [26]:

```
#4) Define a question that would utilize Bernoulli's Equation and perform a
##calculation to support your question.
#my question will be: What is the probability that 3% of flights per day
# from DEN arrive late at their destination?
prob_den = prob_delay_airport['DEN']
prob_ord = prob_delay_airport['ORD']
prob_atl = prob_delay_airport['ATL']


avg_daily = round(data_df.groupby(['month','day']).size().mean())
avg_daily
```

Out[26]:

```
15705
```

In [27]:

```
pmf_den = stats.binom.pmf(round(0.03*avg_daily), n=avg_daily, p=prob_den)
pmf_den
```

Loading [MathJax]/extensions/Safe.js

Out[27]:

```
e-40
```

In [28]:
```python
#therefore, there is a veeeeery small chance that 3% of flights per day are late from DEN
#now, what about just 0.5% of flights?
pmf_den = stats.binom.pmf(round(0.005*avg_daily), n=avg_daily, p=prob_den)
pmf_den
```
Out[28]:
```
7.288349504757996e-35
```
In [29]:
```python
pmf_ord = stats.binom.pmf(round(0.03*avg_daily), n=avg_daily, p=prob_ord)
pmf_ord
```
Out[29]:
```
1.5080216339591587e-12
```
In [30]:
```python
pmf_ord = stats.binom.pmf(round(0.005*avg_daily), n=avg_daily, p=prob_ord)
pmf_ord
```
Out[30]:
```
3.7076640273336703e-66
```
In [31]:
```python
pmf_atl = stats.binom.pmf(round(0.03*avg_daily), n=avg_daily, p=prob_atl)
pmf_atl
```
Out[31]:
```
9.801856178088144e-12
```
In [32]:
```python
pmf_atl = stats.binom.pmf(round(0.005*avg_daily), n=avg_daily, p=prob_atl)
pmf_atl
```
Out[32]:
```
8.169095244079916e-68
```
In [33]:
```python
og_airport_lateCounts[og_airport_lateCounts.origin_airport == 'DEN']
```
Out[33]:

	origin_airport	not_late	late	total	late_prop
85	DEN	113062	80870	193932	0.417002

In [34]:
```python
og_airport_lateCounts[og_airport_lateCounts.origin_airport == 'ORD']
```
Out[34]:

	origin_airport	not_late	late	total	late_prop
		163923	113413	277336	0.408937

Loading [MathJax]/extensions/Safe.js

In [35]:

```
og_airport_lateCounts[og_airport_lateCounts.origin_airport == 'ATL']
```

Out[35]:

	origin_airport	not_late	late	total	late_prop
20	ATL	229245	115034	344279	0.33413

5) Provide a summary of all the values that you calculated for 3 airports

Compare the three to each other.

Which airport would you prefer to fly out of based on your results.

I compared the airports of Denver, Chicago, and Atlanta. I found the following from my calculations:

Denver Airport (DEN):

- 1.541707% probability of delayed arrival.
- mean delay: 34.795994
- std delay: 76.862392
- From Denver, 80870 of 193932 flights had a delayed arrival, a proportion of 0.417002.
- Using Bernoulli's equation, the probability of 3% of flights arriving delayed from DEN was 6.62e-40.
- Using Bernoulli's equation, the probability of 0.5% of flights arriving delayed from DEN was 7.29e-35.

Chicago Airport (ORD):

- 2.162107% probability of delayed arrival.
- mean delay: 40.924092
- std delay: 74.904126
- From Chicago, 113413 of 277336 flights had a delayed arrival, a proportion of 0.408937.
- Using Bernoulli's equation, the probability of 3% of flights arriving delayed from ORD was 1.51e-12.
- Using Bernoulli's equation, the probability of 0.5% of flights arriving delayed from ORD was 3.71e-66.

Atlanta Airport (ATL):

- 2.193010% probability of delayed arrival.
- mean delay: 32.920658
- std delay: 76.114586
- From Atlanta, 115034 of 344279 flights had a delayed arrival, a proportion of 0.33413.
- 's equation, the probability of 3% of flights arriving delayed from ATL was 9.80e-12.

- Using Bernoulli's equation, the probability of 0.5% of flights arriving delayed from ATL was 8.17e-68.

This was a difficult comparison! All 3 of these airports have a high number of delays. Atlanta had the lowest proportion of actual delayed arrivals, while Denver had the lowest probability of delayed arrivals. Atlanta also had the lowest average time delay per delayed flight. Given all of the above, I definitely would not choose the Chicago airport. Between Denver and Atlanta however, I think I would choose the Atlanta airport because I trust the proportion more, the actual data, and Atlanta has the lowest proportion of delayed arrivals.

V. Conclusion

This assignment was interesting as it dealt with real airport data! My parents fly a lot so I shared some of the findings from here with them. They were really interested at seeing the dataset, and weren't surprised at all to see some of those airports as the most delayed. I was a bit confused at exactly what the Bernoulli's equation probabilities were telling me as the probabilities found using the equation were extremely low. This didn't seem realistic, but then I thought that finding the probability of a precise percentage of flights delayed may be a very low probability indeed. Thank you! Please let me know if you have any questions.

All the best, Jeremy

VI. References

MSDS 650 - Week 2 Content:

1.) Class datasets provided for this assignment: flights_clean.csv

2.) From the Experts PDF: Week 2

3.) Sample Assignment (Jupyter Notebook) provided by Professor Hayes

In [ ]:

```

```