Lecture 11: Describing Relationships

Problem Set for Lecture 11

These questions will assist in bolstering your understanding of the material in Lecture 11. Emphasis should be placed on having the correct code/output as well as communication. The deliverable should be a knitted RMarkdown document to pdf without any code running off the page. You will be able to present these in office hours as an oral assessment interview along with the problem set for Lecture 12 and Lecture 13.

***** First, make sure you have read the Lecture Material's writeup *****

1. Using the `Batting` and `People` datasets in the `Lahman` package, create a new dataset called `baseball` by joining the two datasets together and keeping the variables needed for this problem. Then restrict your data to seasons from 1920 or later and to players with at least 100 at-bats in a season. Use this dataset to investigate how offensive performance is related to age, decade, and birthplace. You can remove any missing values that are present. Then answer the following questions.

   (a) Determine the players with the largest total home run counts in this filtered dataset. Display the player ID, first name, last name, and total number of home runs, and identify the top 10.

   (b) For each age, determine the total number of observations and the overall batting average, where batting average is defined by total hits (H) divided by total at-bats (AB). Restrict your attention to ages with at least 10 observations.

   (c) Using the summarized data from the previous part, you might be interested in investigating the relationship between age and batting average. You calculate the correlation as -0.0114 and notice that the visualization shows a non-linear relationship. Because of that you decide to compare the relationship for players less than 30 and those 30 years or older by creating the visualization (found below).

   (d) Create an `age_group` variable with categories `under_25`, `25_29`, `30_34`, as well as `35_or_older`. Also create a `decade` variable from `yearID`. Then create the visualization (found below) that compares the distribution of batting averages across decades for the different age groups.

   (e) Create a summary table showing the mean batting average for each age-group within each decade, where each decade is a row and each age group is its own column.

   (f) Determine the five most common birth states in the dataset. Then, using only those five states, investigate whether birth state and decade appear to be associated. Use counts and conditional proportions to recreate the visualization (found below) to support your conclusion.

2. Using the `Birthdays` dataset in the `mosaicData` package, investigate how birth counts vary across the calendar and across states. Then answer the following questions.

   (a) Determine the average number of births for each day of the week. Which day has the largest average number of births and which has the smallest? (Sunday is the least, Tuesday is the most)

   (b) Compute the mean number of births for each year-month-day-of-week combination. Create the visualization (found below) that compares these mean birth counts across the days of the week.

   (c) Create a table showing the total number of births for each combination of month and day of the week.

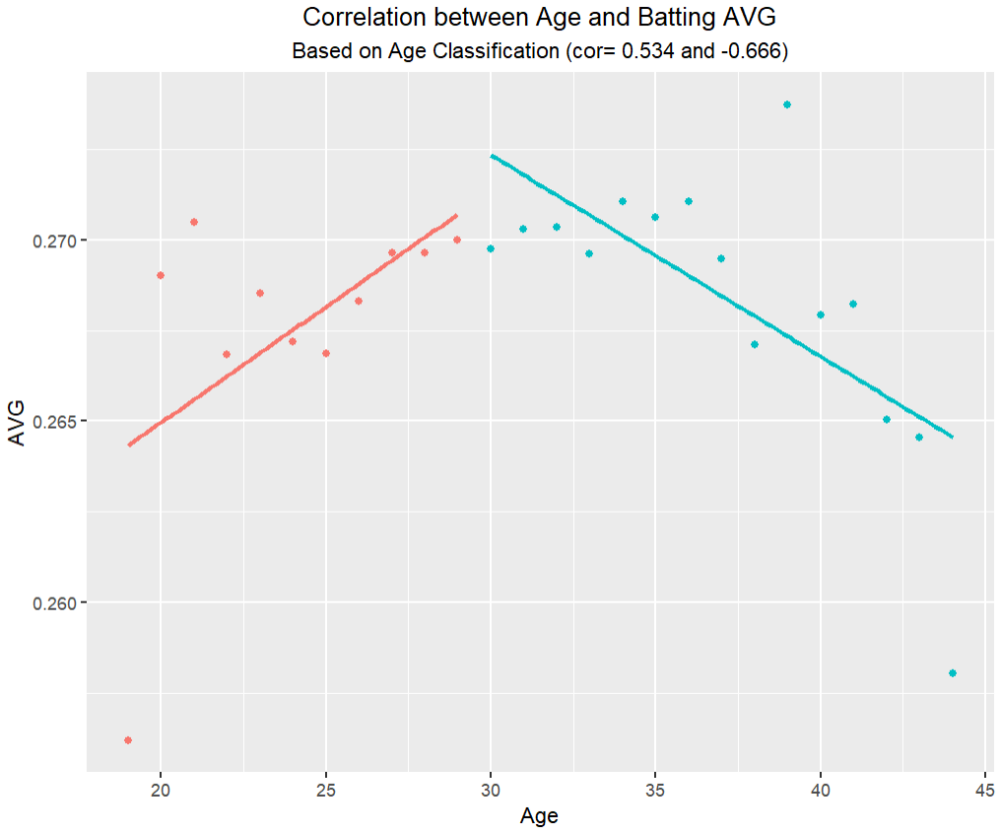
   (d) Determine the percentage of all births that occur on each month-day combination. Then create the visualization (found below) that displays these percentages across the calendar. Within the `scale_fill_steps2()` function, use the following layer within the ggplot plot creation:

```
scale_fill_steps2(low = "red",
                  mid = "white",
                  high = "green",
                  midpoint = 0.275,
                  limits = c(0.22, 0.30),
                  oob = scales::squish,
                  breaks = seq(0.22, 0.30, by = 0.01))
```

   (e) For each state, determine the total number of births in 1970 and in 1988. Use this information to study the relationship between births in 1970 and births in 1988 across states. Create the visualization (found below), calculate the correlation, and describe the relationship in terms of direction, form, strength, and unusual features.

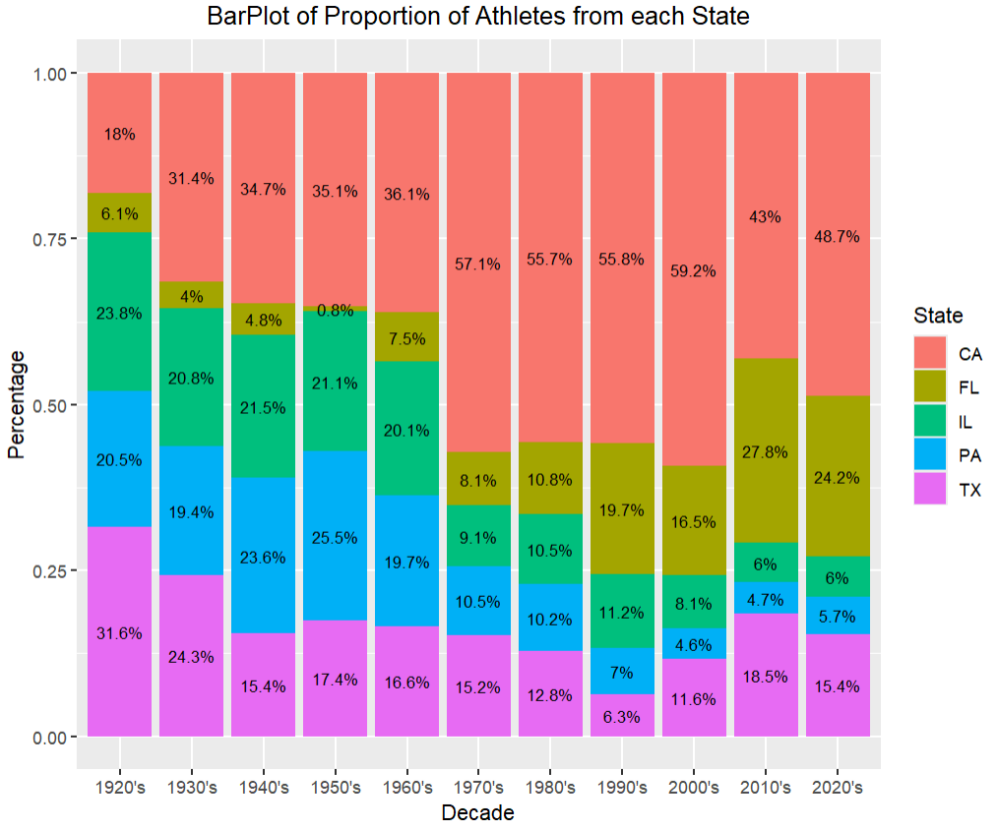
   (f) Using only the locations `DC`, `PA`, `MD`, and `VA`, determine, for each year, the proportion of births contributed by each location out of the total births among those four locations. Then create the visualization (found below) that shows how these proportions change over time.

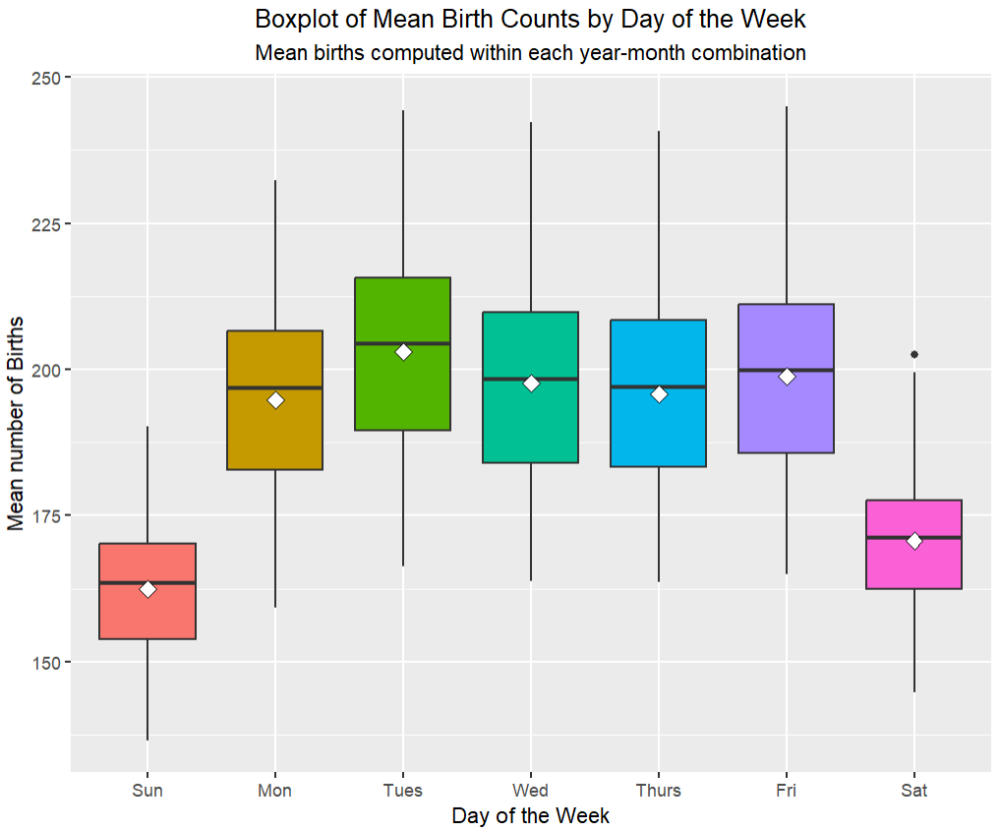
Visualizations and Tables below to Assist you


Correlation between Age and Batting AVG
Based on Age Classification (cor= 0.534 and -0.666)
AVG
0.270
0.265
0.260
20
25
30
35
40
45
Age

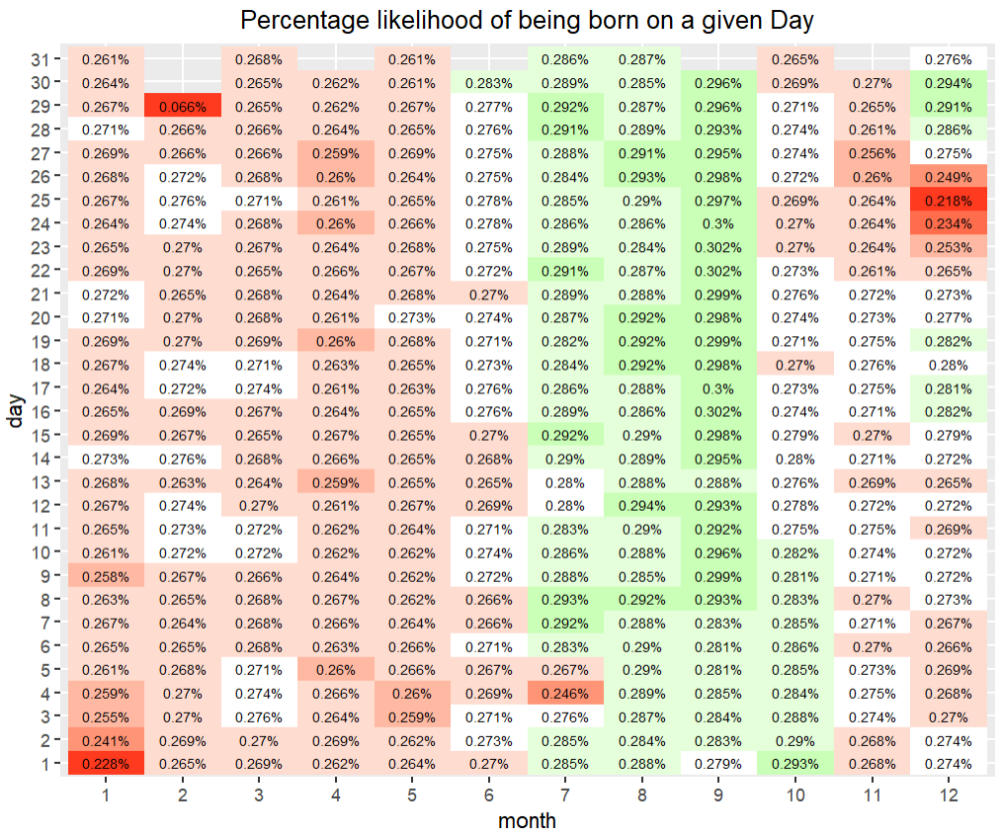

Boxplot of Batting AVG for each Decade
Based on the Age Group
Age Group
under_25
25_29
30_34
35_or_older
AVG
0.4
0.3
0.2
0.1
1920's
1930's
1940's
1950's
1960's
1970's
1980's
1990's
2000's
2010's
2020's
Decade

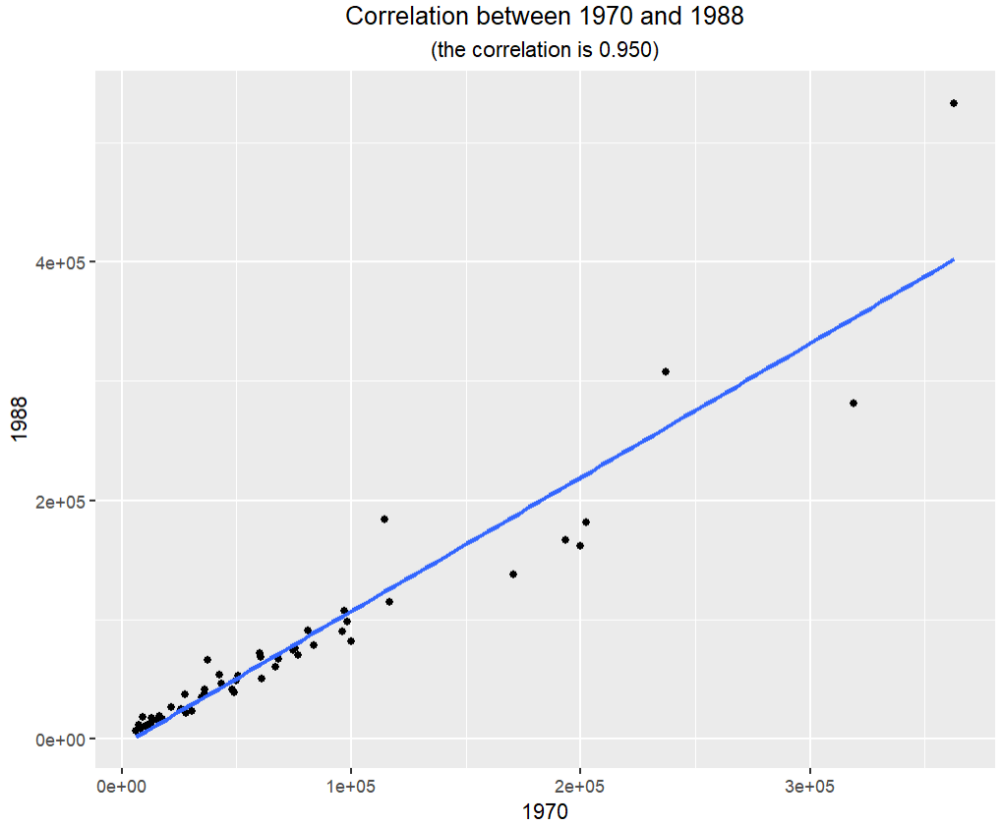
	decade	under_25	‘25_29‘	‘30_34‘	‘35_or_older‘
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	1920’s	0.283	0.287	0.285	0.287
2	1930’s	0.284	0.282	0.284	0.284
3	1940’s	0.266	0.266	0.266	0.272
4	1950’s	0.257	0.262	0.263	0.265
5	1960’s	0.249	0.251	0.250	0.255
6	1970’s	0.253	0.256	0.256	0.257
7	1980’s	0.255	0.258	0.262	0.255
8	1990’s	0.259	0.263	0.265	0.268
9	2000’s	0.265	0.265	0.265	0.264
10	2010’s	0.256	0.253	0.253	0.251
11	2020’s	0.242	0.242	0.244	0.241

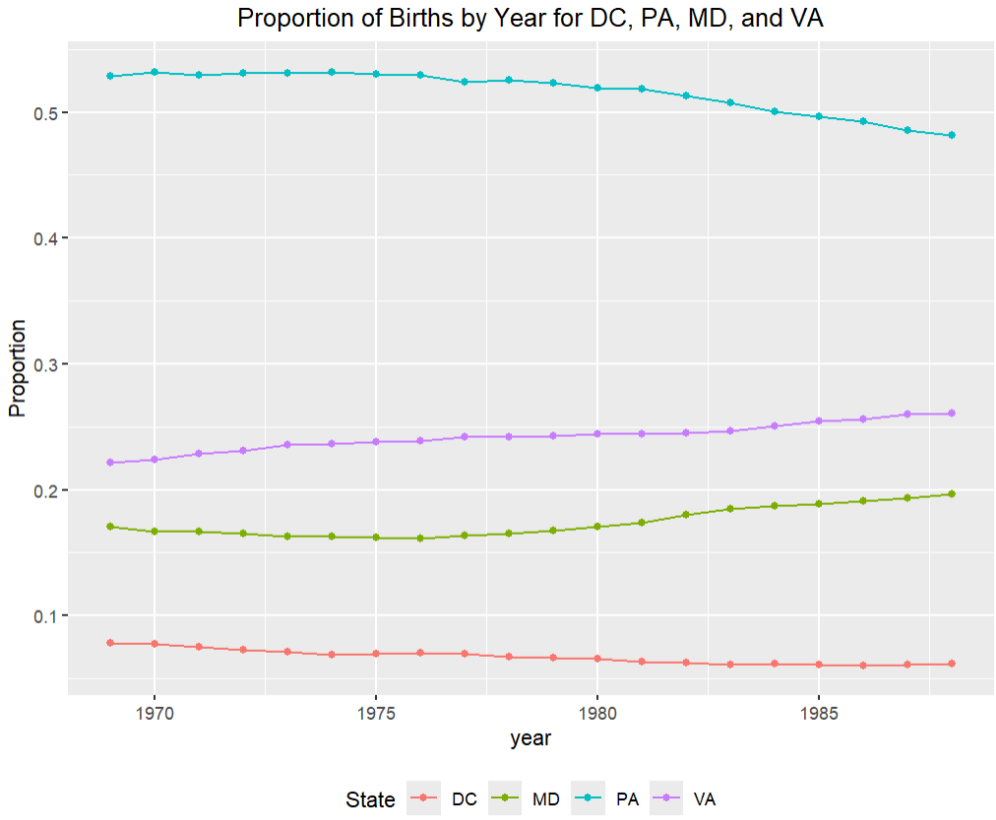




	month	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	1	714590	836139	857521	845615	861021	878440	765839
2	2	666610	791305	818203	799951	786359	807704	692814
3	3	723811	883247	904828	867897	863051	864440	760866
4	4	665550	815348	860751	843635	830436	836368	708815
5	5	711177	830672	885864	851412	866788	882860	756447
6	6	703844	864812	884425	864275	850923	858614	731803
7	7	753943	889500	960908	937586	944654	938384	795320
8	8	786658	935981	961855	934698	917682	952396	820494
9	9	758555	891032	961394	937372	930827	932728	794265
10	10	742663	885689	906782	905599	903144	927794	783701
11	11	715932	863377	888776	845034	802886	843598	735885
12	12	703817	884917	922621	900465	877195	869998	745363


Percentage likelihood of being born on a given Day
day
month
31 0.261% 0.268% 0.261% 0.286% 0.287% 0.265% 0.276%
30 0.264% 0.265% 0.262% 0.261% 0.283% 0.289% 0.285% 0.296% 0.269% 0.27% 0.294%
29 0.267% 0.066% 0.265% 0.262% 0.267% 0.277% 0.292% 0.287% 0.296% 0.271% 0.265% 0.291%
28 0.271% 0.266% 0.266% 0.264% 0.265% 0.276% 0.291% 0.289% 0.293% 0.274% 0.261% 0.286%
27 0.269% 0.266% 0.266% 0.259% 0.269% 0.275% 0.288% 0.291% 0.295% 0.274% 0.256% 0.275%
26 0.268% 0.272% 0.268% 0.26% 0.264% 0.275% 0.284% 0.293% 0.298% 0.272% 0.26% 0.249%
25 0.267% 0.276% 0.271% 0.261% 0.265% 0.278% 0.285% 0.29% 0.297% 0.269% 0.264% 0.218%
24 0.264% 0.274% 0.268% 0.26% 0.266% 0.278% 0.286% 0.286% 0.3% 0.27% 0.264% 0.234%
23 0.265% 0.27% 0.267% 0.264% 0.268% 0.275% 0.289% 0.284% 0.302% 0.27% 0.264% 0.253%
22 0.269% 0.27% 0.265% 0.266% 0.267% 0.272% 0.291% 0.287% 0.302% 0.273% 0.261% 0.265%
21 0.272% 0.265% 0.268% 0.264% 0.268% 0.27% 0.289% 0.288% 0.299% 0.276% 0.272% 0.273%
20 0.271% 0.27% 0.268% 0.261% 0.273% 0.274% 0.287% 0.292% 0.298% 0.274% 0.273% 0.277%
19 0.269% 0.27% 0.269% 0.26% 0.268% 0.271% 0.282% 0.292% 0.299% 0.271% 0.275% 0.282%
18 0.267% 0.274% 0.271% 0.263% 0.265% 0.273% 0.284% 0.292% 0.298% 0.27% 0.276% 0.28%
17 0.264% 0.272% 0.274% 0.261% 0.263% 0.276% 0.286% 0.288% 0.3% 0.273% 0.275% 0.281%
16 0.265% 0.269% 0.267% 0.264% 0.265% 0.276% 0.289% 0.286% 0.302% 0.274% 0.271% 0.282%
15 0.269% 0.267% 0.265% 0.267% 0.265% 0.27% 0.292% 0.29% 0.298% 0.279% 0.27% 0.279%
14 0.273% 0.276% 0.268% 0.266% 0.265% 0.268% 0.29% 0.289% 0.295% 0.28% 0.271% 0.272%
13 0.268% 0.263% 0.264% 0.259% 0.265% 0.265% 0.28% 0.288% 0.288% 0.276% 0.269% 0.265%
12 0.267% 0.274% 0.27% 0.261% 0.267% 0.269% 0.28% 0.294% 0.293% 0.278% 0.272% 0.272%
11 0.265% 0.273% 0.272% 0.262% 0.264% 0.271% 0.283% 0.29% 0.292% 0.275% 0.275% 0.269%
10 0.261% 0.272% 0.272% 0.262% 0.262% 0.274% 0.286% 0.288% 0.296% 0.282% 0.274% 0.272%
9 0.258% 0.267% 0.266% 0.264% 0.262% 0.272% 0.288% 0.285% 0.299% 0.281% 0.271% 0.272%
8 0.263% 0.265% 0.268% 0.267% 0.262% 0.266% 0.293% 0.292% 0.293% 0.283% 0.27% 0.273%
7 0.267% 0.264% 0.268% 0.266% 0.264% 0.266% 0.292% 0.288% 0.283% 0.285% 0.271% 0.267%
6 0.265% 0.265% 0.268% 0.263% 0.266% 0.271% 0.283% 0.29% 0.281% 0.286% 0.27% 0.266%
5 0.261% 0.268% 0.271% 0.26% 0.266% 0.267% 0.267% 0.29% 0.281% 0.285% 0.273% 0.269%
4 0.259% 0.27% 0.274% 0.266% 0.26% 0.269% 0.246% 0.289% 0.285% 0.284% 0.275% 0.268%
3 0.255% 0.27% 0.276% 0.264% 0.259% 0.271% 0.276% 0.287% 0.284% 0.288% 0.274% 0.27%
2 0.241% 0.269% 0.27% 0.269% 0.262% 0.273% 0.285% 0.284% 0.283% 0.29% 0.268% 0.274%
1 0.228% 0.265% 0.269% 0.262% 0.264% 0.27% 0.285% 0.288% 0.279% 0.293% 0.268% 0.274%
1 2 3 4 5 6 7 8 9 10 11 12


Correlation between 1970 and 1988
(the correlation is 0.950)
1988
1970
4e+05
2e+05
0e+00
0e+00
1e+05
2e+05
3e+05


Proportion of Births by Year for DC, PA, MD, and VA
Proportion
0.5
0.4
0.3
0.2
0.1
1970
1975
1980
1985
year
State
DC
MD
PA
VA