

School of Computer Science

Data Wrangling in Fulfilment of

DATA9910

Maksymilian Drzezdzon

C15311966

Degree: TU060/1

Module Coordinator: Brendan Tierney

Declaration of Ownership: I declare that the attached work is entirely my own and that all sources have been acknowledged.

**Date: 2020/11/25**

Section A – Data Analysis

This section will focus on examining a medical insurance dataset of 1338 rows and 7 columns.[1] Collected variables include age, sex, BMI, number of children, does the person smoke, region of residence within the US and their annual insurance charges.

The main question for this part of the analysis is “what factors affect medical insurance costs the most”. With a limited amount of time for each section only a few areas will be explored to provide some insight meanwhile showcasing SQLs analytical functions with hypothesis testing, correlations and descriptive analytics etc.

Results for each query will be added within the code snippet for single row and column results using a /\* result here\*/ or with another snippet for multi row and column results, adding limit 1 to a query that only returns a single value is redundant.

A five-number summary is a set of descriptive statistics used to analyse a dataset; this was completed on each column of numeric data. A five-number summary consists of the min, max, median and first and third quartiles of data, a template would look like:

The next five queries are part of the five-number summary and are run against each numeric column to get a better understanding of the dataset, the column below is the charges column, representing annual medical insurance charges.

select round(max(charges),2) "max charges cost" from insurance; /\* 63,770.43 dollars \*/

The function above selects the charges from the insurance table, gets the highest value and rounds it to the 2nd decimal place and slaps on a “max charges cost” label.

select round(min(charges),2) "min charges cost" from insurance; /\* 1,121.87 dollars \*/

The function selects the charges from the insurance table, gets the lowest value and rounds it to the 2nd decimal place and adds a “min charges cost” label.

select median(charges) "middle charges cost value" from insurance; /\* 9,382.03 dollars \*/

The function above selects the charges from the insurance table, gets the median value which separates the second and third quartiles, then adds a “middle charges cost” label.

The above three queries were just used to get an idea of potential outliers.

1. select round(sum(charges), 2)"insurance costs in first quartile" from (select charges, ntile(4) over(order by charges) as QUARTILE from insurance) where quartile = 1; /\* 955,784.96 dollars \*/

2 select round(sum(charges), 2)"insurance costs in third quartile" from (select charges, ntile(4) over(order by charges) as QUARTILE from insurance) where quartile = 3; /\* 4,050,700.59 dollars \*/

The two above queries sum up and round to two decimal places any data that falls into the bottom twenty five percent. This is completed by the nested query that selects the charges column and puts it into four buckets with ntile(4) followed by selecting the charges rows with the over() function and the ordering it by charges from the insurance table. Lastly its tagged as quartile and used to grab the first and third quartile from the four buckets with where quartile is equal to three. An appropriate label is added to each resulting column.

The goal of this query was to get a more tangible idea of how much of the overall cost is under which quartile, if sum() wasn’t used the query would return a list of all records with an ntile number which isn’t interpretable.

3. select STATS\_ONE\_WAY\_ANOVA(bmi, charges, 'F\_RATIO') f\_ratio from insurance; /\*1.11\*/

/\*an f ratio close to 1 means that the null hypothesis is true, meaning there is no correlation between the two\*/

4. select STATS\_ONE\_WAY\_ANOVA(age, charges, 'SIG') p\_value from insurance; /\* 0.0000000000000001 \*/

/\* a p-value of more than 0.05 is not statistically significant meaning there is evidence for age having an impact on health insurance charges\*/

 These two queries where used to calculate the significance toward finding out what values have a statistical significance on insurance costs. However, what was a profound predictor of medical insurance was age and if the person smoked or not, both yielding a p value of close to 0. [3]

update insurance set smoker = 0 where smoker = 'no';

update insurance set smoker = 1 where smoker = 'yes';

5. select STATS\_ONE\_WAY\_ANOVA(smoker, charges, 'SIG') p\_value from insurance;

update insurance set smoker = 'no' where smoker = '0';

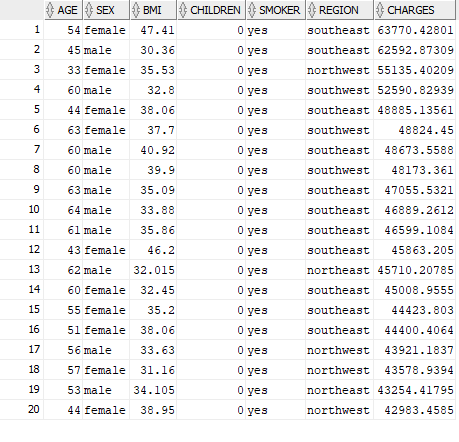
update insurance set smoker = 'yes' where smoker = '1';

A quick test was run to make sure that the values yes and no were being interpreted properly by SQL and were converted to a 1 or 0 for the one-way anova, the result was still statistically significant

6. select STATS\_F\_TEST(smoker, charges) two\_tailed\_test from insurance; /\* 0.0000000000000000000000000000000000000000000000000003 \*/

This is the further illustrated by the two tailed test, a two tailed test ranges from 0 to 1 the closer the result is to 0 the more statistically significant it is that smoking does have an impact on medical charge costs. Alternately the ‘TWO\_SIDED\_SIG’ argument can be used which would yield the same result as a two tailed test is run by default when no test parameter is passed in. [2] The goal here was to confirm findings from the one way anova with a two tailed test.

select \* from insurance order by charges desc fetch next 20 rows only;



The top 20 highest medical insurance charges are all middle aged or older smokers, with no kids. The goal was to see if the above statistics are true and what parameters each individual would have when paying a premium on insurance.

The query selects all rows from the insurance table and orders it by the charges column limiting it to 20 rows, originally the query had children in it but as all individuals have no kids that variable was then removed as its redundant.

7. select PERCENT\_RANK(50, 'yes') within group

(ORDER BY age, smoker)

from insurance; /\*0.73\*/

7.5 select PERCENT\_RANK(50, 'no') within group

(ORDER BY age, smoker)

from insurance; /\*0.71\*/

The goal of the above query was to see the significance of smoking on an individual ranking trying this out across a range of ages its usually 1-3%. The query takes a percent rank of an individual at the age of 50 that smokes from the group ordered by age and smoke columns from the insurance table.

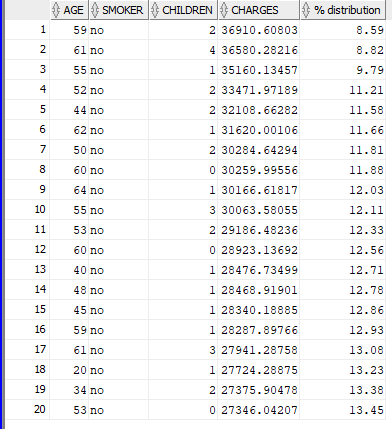
8. select age, smoker, children, charges,

round(cume\_dist() over (order by charges desc)\*100,2)"% distribution"

from insurance

order by smoker asc, "% distribution" asc

fetch next 20 rows only;



The above query grabs a cumulative distribution of individuals by their medical charges, smoking habits and % of distribution in descending order. The goal here was to put things together and validate previous findings.

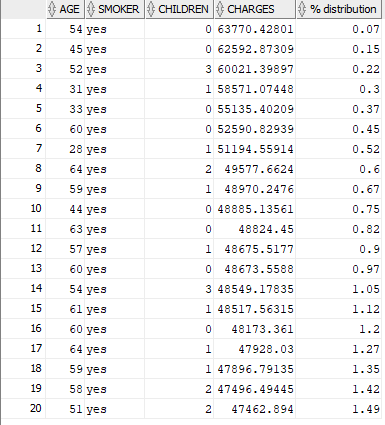
1. select age, smoker, children, charges,

round(cume\_dist() over (order by charges desc)\*100,2)"% distribution"

from insurance

order by charges desc, "% distribution" asc

fetch next 20 rows only;



When results are sorted by charges rather than whether someone is a smoker or not, the difference in medical charges almost doubles in some cases. Individuals in these groups cover a low amount of the overall however

1. select age, smoker, children, charges,

round(cume\_dist() over (order by charges desc)\*100,2)"% distribution"

from insurance

where (age = 18 and smoker = 'yes')

order by charges desc

fetch next 20 rows only;

select age, smoker, children, charges,

round(cume\_dist() over (order by charges desc)\*100,2)"% distribution"

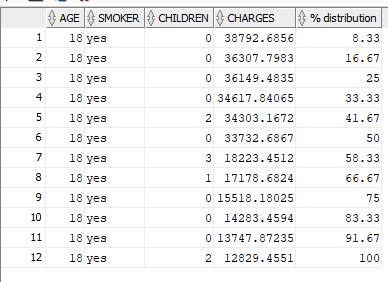
from insurance

where (age = 18 and smoker = 'no')

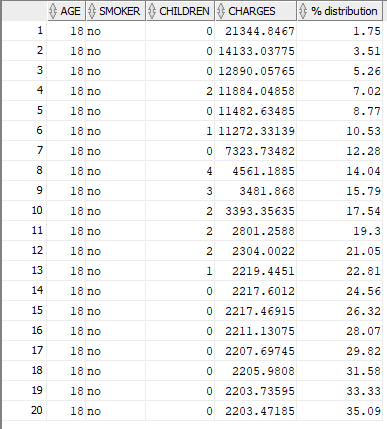
order by charges desc

fetch next 20 rows only;

when compared to eighteen-year-olds that smoke, which there are fewer categories and a much higher distribution.



When compared to the same group but instead that dont smoke, its much more spread out with lower distributions. The goal of these last few queries was to get an idea of how many people are paying more for insurance charges and have unhealthy habits, smoking and BMI were used to asses that. Maybe a campaign that offered younger customers health insurance discounts if they didn’t smoke could incentivise people to look after their health and save money. This might not be at the benefit of the insurance company but it’s an observation nonetheless.



Section B – Data Audit Report

Section C – Machine Learning

Appendix

Oracle 1996, 2007 <https://docs.oracle.com/database/121/SQLRF/functions186.htm#SQLRF06318> [2]

<https://docs.oracle.com/database/121/SQLRF/functions190.htm#SQLRF06322> [3]

https://www.oracletutorial.com/oracle-analytic-functions/

References

Miri Choi, 2017, [https://www.kaggle.com/mirichoi0218/insurance [1](https://www.kaggle.com/mirichoi0218/insurance%20%5b1)]