Certification Program on Business Analytics

Assignment No:2

Date: 18th Feb to 24th Feb-2018

Submitted By:
Mallesham Yamulla

Topics

1. Read and Understand the TWO Papers titled
   (i) “Exploratory Data Analysis”
   (ii) “Analytics and Data Science”

Make a brief Summary on the main theme of each paper.

2. Perform Descriptive Analytics on the attached “Height Change in a Day” Dataset (corrected) and give your findings.
3. Perform Descriptive Analytics on the attached “Crime Statistics” Dataset and give your findings
4. Data visualization on Global superstore dataset using Tableau/Power BI

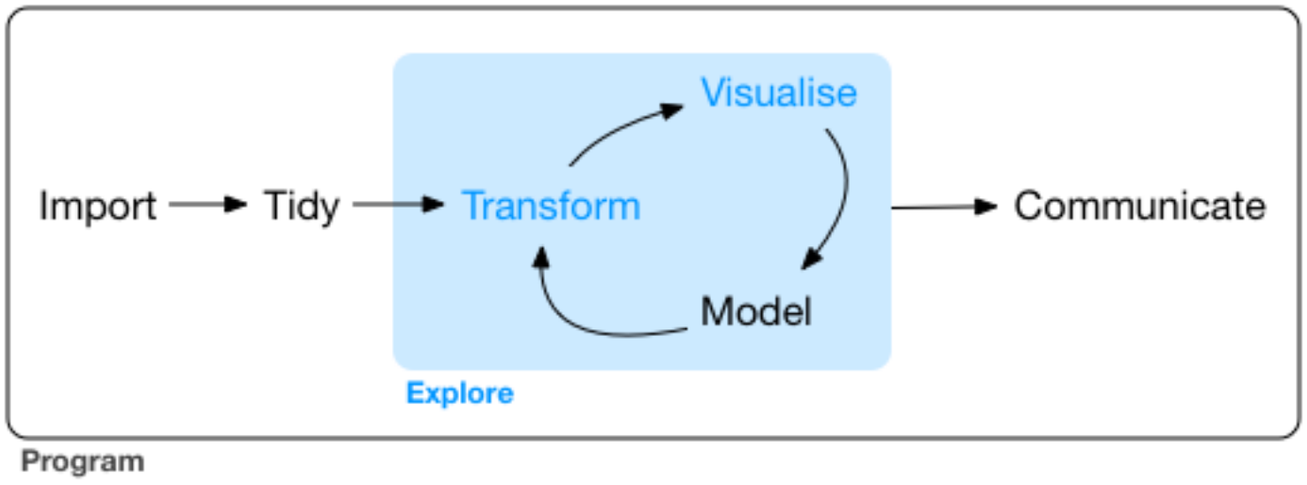
Exploratory Data Analysis

Exploratory Data Analysis (EDA) is used on the one hand to answer questions, test business assumptions, generate hypotheses for further analysis. On the other hand, we can also use it to prepare the data for modeling. The thing that these two probably have in common is a good knowledge of our data to either get the answers that we need or to develop an intuition for interpreting the results of future modeling.

There are a lot of ways to reach these goals: we can get a basic description of the data, visualise it, identify patterns in it, identify challenges of using the data, etc.

One of the things that we will often see when w're reading about EDA is Data profiling. Data profiling is concerned with summarising our dataset through descriptive statistics. We want to use a variety of measurements to better understand our dataset.

The goal of data profiling is to have a solid understanding of our data so we can afterwards start querying and visualising our data in various ways. However, this doesn't mean that we don't have to iterate: exactly because data profiling is concerned with summarising our dataset, it is frequently used to assess the data quality. Depending on the result of the data profiling, we might decide to correct, discard or handle our data differently



Analytics and Data Science

Predictive Analytics:

Predictive analytics consists of techniques that uses models constructed from the past data to predict the future or ascertain the impact of one variable on another.

Example:

1. Past data on the sales might be used to construct a model to predict future sales which can factor in the product's growth trajectory and seasonality based on past patterns.

2. Survey data and purchase behavious might also be useful to predict the market for a new product.

Predictive analytics is data science, a multidisciplinary skill set essential for success in business,nonprofit organizations, and government

Data science teams will spend most of their time collecting and storing data, and then using that data to ask questions.They create reports using statistics and math to see if they can get at answers.

To be in a data science team we need to be familiar with as we explore statistical analysis, these are:

1. **Descriptive statistics:** The process of analyzing, describing or summarizing data in a meaningful way to discover patterns in the data

2. **Predicive analytics:** Applying statistical analysis to historical data in an effort to predict the future

3. **Probability:** The likelihood that something will happen

4. **Correlation:** A series of statistical relationaships that measures the degree ti which the things are related. It's ususally measured as number between 1 or 0

5. **Causation:** When one event is the result of other occurance of another event.

Analytics and Data Science

And again going back to Predictive analytics, it involves searching for meaningful relationships among variables and representing those relationships in models. These are as follows,

1. Response variables—things we are trying to predict.
2. Explanatory variables or predictors—things we observe, manip-ulate, or control that could relate to the response

Predictive modeling can be carried out by Regression and classification models.

Regression:

Regression predictive modeling is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y).

A continuous output variable is a real-value, such as an integer or floating point value. These are often quantities, such as amounts and sizes.

For example, a house may be predicted to sell for a specific dollar value, perhaps in the range of $100,000 to $200,000.

- A regression problem requires the prediction of a quantity.
- A regression can have real valued or discrete input variables.
- A problem with multiple input variables is often called a multivariate regression problem.
- A regression problem where input variables are ordered by time is called a time series forecasting problem

Analytics and Data Science

Classification:

Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation.

For example, an email of text can be classified as belonging to one of two classes: "spam" and "*not spam*".

- A classification problem requires that examples be classified into one of two or more classes.
- A classification can have real-valued or discrete input variables.
- A problem with two classes is often called a two-class or binary classification problem.
- A problem with more than two classes is often called a multi-class classification problem.
- A problem where an example is assigned multiple classes is called a multi-label classification problem.

Analytics and Data Science

We should consider the below three general approaches to research and modeling in predictive analytics.

Traditional : The traditional approach to research and modeling begins with the specification of a theory or model

Data- Adaptive : It starts with data, then look through it find useful predictors and we should give little thought to the theories or hypothesis prior to running analysis.

Model dependent: It begins with the spec- ification of a model and uses that model to generate data, predictions, or recommendations. Simulations and mathematical programming methods, primary tools of operations research, are examples of model-dependent research.

We may employ classical or Bayesian methods to make predictions and Our approach to predictive analytics is based upon a simple premise: **The value of a model lies in the quality of its predictions**

What is training data?

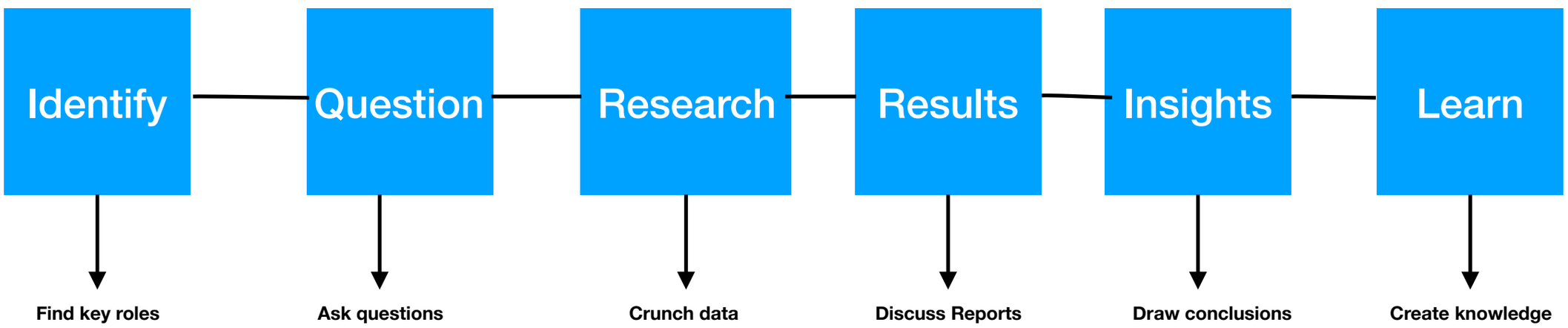
Simply put, training data is used to train an algorithm. Generally, training data is a certain percentage of an overall dataset along with testing set. As a rule, the better the training data, the better the algorithm or classifier performs.

What is a test set?

Once a model is trained on a training set, it's usually evaluated on a test set. Oftentimes, these sets are taken from the same overall dataset, though the training set should be labeled or enriched to increase an algorithm's confidence and accuracy.

Analytics and Data Science

For Data science project we can use the below Data science life cycle, it has six steps as shown in the below diagram.



Descriptive analytics on the data of height changes in a day

Introduction: I have carried out the descpritive analytics on the given dataset in order to figure out what factors are invloved on growing a person height in a day using R programming, and I have made use of Tidyverse package to do comphrenshive analysis of the mentioned data

Structure of the dataset:

```
> glimpse(altura_en_dia)
Observations: 30
Variables: 17
$ Stress-Personal     <chr> "L", "L", "M", "M", "M", "L", "L", "L", "M", "L", "M", "H", "H", "M", "L", "M", "M", "L", "H", "M", "M", "M", "M", "L", "L", "M", "M", "M", "L", "L"
$ Stress-Professional <chr> "L", "M", "M", "M", "M", "L", "M", "L", "M", "H", "M", "H", "L", "L", "M", "H", "M", "L", "M", "L", "M", "L", "M", "L", "L", "M", "M", "M", "L", "L"
$ Activity_Level      <chr> "M", "H", "H", "L", "L", "M", "M", "H", "M", "L", "H", "H", "M", "L", "L", "M", "L", "L", "M", "M", "M", "M", "L", "M", "M", "L", "M", "H", "M", "L"
$ Age                 <dbl> 28, 30, 37, 34, 33, 24, 24, 24, 37, 25, 31, 35, 35, 33, 31, 23, 33, 27, 31, 30, 24, 42, 41, 24, 23, 29, 28, 37, 33, 28
$ Sex                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "F", "M", "M", "F", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M"
$ Height_ cm          <dbl> 176.0, 184.5, 166.0, 183.0, 174.0, 172.0, 180.0, 170.0, 167.0, 157.5, 162.0, 166.0, 154.0, 174.5, 174.0, 170.5, 167.0, 178.0, 165.0, 171.0, 181.5, 171.0, 1...
$ Weight_Kg           <dbl> 64.3, 70.0, 74.0, 123.0, 76.2, 73.9, 82.8, 70.2, 78.7, 56.1, 57.6, 72.9, 65.6, 93.0, 90.0, 73.8, 81.9, 79.1, 78.9, 80.9, 102.2, 83.9, 73.5, 57.1, 50.1, 89....
$ Waist_ cm           <dbl> 87, 84, 96, 120, 89, 92, 97, 88, 92, 75, 83, 92, 98, 100, 102, 93, 97, 104, 97, 100, 102, 102, 92, 78, 85, 100, 87, 99, 93, 105
$ BP-Systolic         <dbl> 117, 133, 90, 97, 110, 146, 118, 178, 103, 90, 112, 129, 109, 138, 134, 145, 114, 118, 118, 122, 133, 130, 149, 118, 104, 121, 125, 112, 130, 126
$ BP-Diastolic        <dbl> 78, 82, 60, 63, 73, 86, 73, 64, 60, 66, 77, 74, 69, 83, 86, 70, 89, 70, 81, 82, 80, 70, 91, 69, 62, 72, 74, 67, 69, 77
$ Pulse               <dbl> 111, 103, 65, 91, 81, 98, 80, 84, 99, 72, 92, 77, 100, 95, 99, 89, 109, 106, 69, 110, 71, 99, 96, 101, 99, 89, 92, 68, 71, 103
$ BMI                 <dbl> 20.8, 20.6, 26.9, 36.7, 25.2, 25.0, 25.6, 24.3, 28.2, 22.6, 22.2, 26.5, 27.7, 30.5, 29.7, 25.4, 29.4, 25.0, 29.0, 27.6, 31.0, 28.7, 25.0, 20.8, 22.3, 32.1,...
$ Body-Fat            <dbl> 20.2, 20.1, 25.0, 36.7, 23.9, 24.3, 23.3, 20.4, 29.7, 32.9, 24.1, 28.0, 35.5, 30.2, 29.5, 27.0, 35.4, 27.3, 36.8, 26.9, 30.4, 38.9, 46.5, 76.0, 29.9, 31.0,...
$ Body-Age            <dbl> 28, 29, 47, 67, 42, 37, 39, 33, 52, 34, 33, 47, 50, 54, 52, 39, 54, 42, 51, 46, 52, 60, 56, 18, 29, 53, 38, 52, 51, 47
$ Cal-K               <dbl> 1529.00, 1628.00, 1654.00, 23.48, 1702.00, 1665.00, 1820.00, 1624.00, 1703.00, 1200.00, 1395.00, 1621.00, 1332.00, 191.00, 1890.00, 1649.00, 1722.00, 1734....
$ HI                  <dbl> 81, 57, 34, 43, 19, 66, 32, 63, 25, 53, 55, 28, 68, 38, 25, 32, 20, 37, 17, 63, 42, 23, 27, 51, 58, 55, 32, 39, 33, 9
$ Change              <dbl> 0.6, 3.5, 7.0, 0.3, 0.3, 0.2, 0.6, 0.4, 1.0, 1.0, 4.0, 1.8, 1.3, 0.9, 2.1, 1.0, 0.7, 2.5, 2.0, 2.2, 0.7, 1.5, 0.5, 0.3, 2.0, 1.0, 1.5, 4.3, 1.0, 0.5
> 
```

Descriptive analytics on the data of height changes in a day

As part of data transformations on data, I would like add up the below 3 new variables to the data frame to see out and categorise the group of students based on the available information of their BP-Systolic, BP-Diastolic, BMI Levels and Ages

Blood pressure levels:

BP-Systolic	BP-Diastolic	Blood_pressure_levels
<120	<80	Normal
Between(120,139)	Between(80,89)	Prehyper
Between(120,139)	Between(120,139)	Stage 1 hypertension
>=160	>=100	Stage 2 hypersion

Descriptive analytics on the data of height changes in a day

Age groups:

AGE	Age Group
Between (18,25)	Age1
Between(26,30)	Age2
Between(31,35)	Age3
Between(36,40)	Age4
>=41	Age5

Weight levels:

BMI Level	Weight_levels
<18.5	Underweight
Between(18.5,24.9)	Normal Weight
Between(25,29.5)	Over weight
>=30	Obese

Descriptive analytics on the data of height changes in a day

We can look at this data after perfomring the required transformations

```
> glimpse(altura_en_dia_agregado)
Observations: 30
Variables: 21
$ Stress-Personal       <chr> "L", "L", "M", "M", "M", "L", "L", "L", "M", "L", "M", "H", "H", "M", "L", "M", "M", "L", "H", "M", "M", "M", "M", "L", "L", "M", "M", "M", "L", "L"
$ Stress-Professional   <chr> "L", "M", "M", "M", "M", "L", "M", "L", "M", "H", "M", "H", "L", "L", "M", "H", "M", "L", "M", "L", "M", "L", "M", "L", "L", "M", "M", "M", "L", "L"
$ Activity_Level        <chr> "M", "H", "H", "L", "L", "M", "M", "H", "M", "L", "H", "H", "M", "L", "L", "M", "L", "L", "M", "M", "M", "M", "L", "M", "M", "L", "M", "H", "M", "L"
$ Age                   <dbl> 28, 30, 37, 34, 33, 24, 24, 24, 37, 25, 31, 35, 35, 33, 31, 23, 33, 27, 31, 30, 24, 42, 41, 24, 23, 29, 28, 37, 33, 28
$ Sex                   <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "F", "M", "M", "F", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M", "M"
$ Height_ cm            <dbl> 176.0, 184.5, 166.0, 183.0, 174.0, 172.0, 180.0, 170.0, 167.0, 157.5, 162.0, 166.0, 154.0, 174.5, 174.0, 170.5, 167.0, 178.0, 165.0, 171.0, 181.5, 171.0,...
$ Weight_Kg             <dbl> 64.3, 70.0, 74.0, 123.0, 76.2, 73.9, 82.8, 70.2, 78.7, 56.1, 57.6, 72.9, 65.6, 93.0, 90.0, 73.8, 81.9, 79.1, 78.9, 80.9, 102.2, 83.9, 73.5, 57.1, 50.1, 8...
$ Waist_ cm             <dbl> 87, 84, 96, 120, 89, 92, 97, 88, 92, 75, 83, 92, 98, 100, 102, 93, 97, 104, 97, 100, 102, 102, 92, 78, 85, 100, 87, 99, 93, 105
$ BP-Systolic           <dbl> 117, 133, 90, 97, 110, 146, 118, 178, 103, 90, 112, 129, 109, 138, 134, 145, 114, 118, 118, 122, 133, 130, 149, 118, 104, 121, 125, 112, 130, 126
$ BP-Diastolic          <dbl> 78, 82, 60, 63, 73, 86, 73, 64, 60, 66, 77, 74, 69, 83, 86, 70, 89, 70, 81, 82, 80, 70, 91, 69, 62, 72, 74, 67, 69, 77
$ Pulse                 <dbl> 111, 103, 65, 91, 81, 98, 80, 84, 99, 72, 92, 77, 100, 95, 99, 89, 109, 106, 69, 110, 71, 99, 96, 101, 99, 89, 92, 68, 71, 103
$ BMI                   <dbl> 20.8, 20.6, 26.9, 36.7, 25.2, 25.0, 25.6, 24.3, 28.2, 22.6, 22.2, 26.5, 27.7, 30.5, 29.7, 25.4, 29.4, 25.0, 29.0, 27.6, 31.0, 28.7, 25.0, 20.8, 22.3, 32....
$ Body-Fat              <dbl> 20.2, 20.1, 25.0, 36.7, 23.9, 24.3, 23.3, 20.4, 29.7, 32.9, 24.1, 28.0, 35.5, 30.2, 29.5, 27.0, 35.4, 27.3, 36.8, 26.9, 30.4, 38.9, 46.5, 76.0, 29.9, 31....
$ Body-Age              <dbl> 28, 29, 47, 67, 42, 37, 39, 33, 52, 34, 33, 47, 50, 54, 52, 39, 54, 42, 51, 46, 52, 60, 56, 18, 29, 53, 38, 52, 51, 47
$ Cal-K                 <dbl> 1529.00, 1628.00, 1654.00, 23.48, 1702.00, 1665.00, 1820.00, 1624.00, 1703.00, 1200.00, 1395.00, 1621.00, 1332.00, 191.00, 1890.00, 1649.00, 1722.00, 173...
$ HI                    <dbl> 81, 57, 34, 43, 19, 66, 32, 63, 25, 53, 55, 28, 68, 38, 25, 32, 20, 37, 17, 63, 42, 23, 27, 51, 58, 55, 32, 39, 33, 9
$ Change                <dbl> 0.6, 3.5, 7.0, 0.3, 0.3, 0.2, 0.6, 0.4, 1.0, 1.0, 4.0, 1.8, 1.3, 0.9, 2.1, 1.0, 0.7, 2.5, 2.0, 2.2, 0.7, 1.5, 0.5, 0.3, 2.0, 1.0, 1.5, 4.3, 1.0, 0.5
$ height_de_cambiado    <dbl> 176.6, 188.0, 173.0, 183.3, 174.3, 172.2, 180.6, 170.4, 168.0, 158.5, 166.0, 167.8, 155.3, 175.4, 176.1, 171.5, 167.7, 180.5, 167.0, 173.2, 182.2, 172.5,...
$ age_grupo             <chr> "Age2", "Age2", "Age4", "Age3", "Age3", "Age1", "Age1", "Age1", "Age4", "Age1", "Age3", "Age3", "Age3", "Age3", "Age3", "Age1", "Age3", "Age2", "Age3", "...
$ Weight_levels         <chr> "Normal Weight", "Normal Weight", "Over Weight", "Obese", "Over Weight", "Over Weight", "Over Weight", "Normal Weight", "Over Weight", "Normal Weight", "...
$ Blood_pressure_levels <chr> "Normal", "Prehyper", "Normal", "Normal", "Normal", "Prehyper", "Normal", "Stage 2 hypertension", "Normal", "Normal", "Normal", "Prehyper", "Normal", "Pr...
>
```

Summary stats:

```
> summary(altura_reporte)
      Age          Height_ cm         BMI              HI           Change      Height_after_change_cm
 Min.   :23.00   Min.   :150.0   Min.   :20.60   Min.   : 9.00   Min.   :0.200   Min.   :152.0
 1st Qu.:25.50   1st Qu.:166.0   1st Qu.:25.00   1st Qu.:27.25   1st Qu.:0.600   1st Qu.:167.6
 Median :30.50   Median :170.2   Median :26.70   Median :37.50   Median :1.000   Median :171.8
 Mean   :30.47   Mean   :169.6   Mean   :26.65   Mean   :40.83   Mean   :1.557   Mean   :171.2
 3rd Qu.:33.75   3rd Qu.:174.0   3rd Qu.:28.93   3rd Qu.:55.00   3rd Qu.:2.000   3rd Qu.:175.1
 Max.   :42.00   Max.   :184.5   Max.   :36.70   Max.   :81.00   Max.   :7.000   Max.   :188.0
>
```

Descriptive analytics on the data of height changes in a day

Formulate questions:

1. How are the height growths at different ages, is it grwing higher in group of Age5 and in which of age groups it's growing higher?

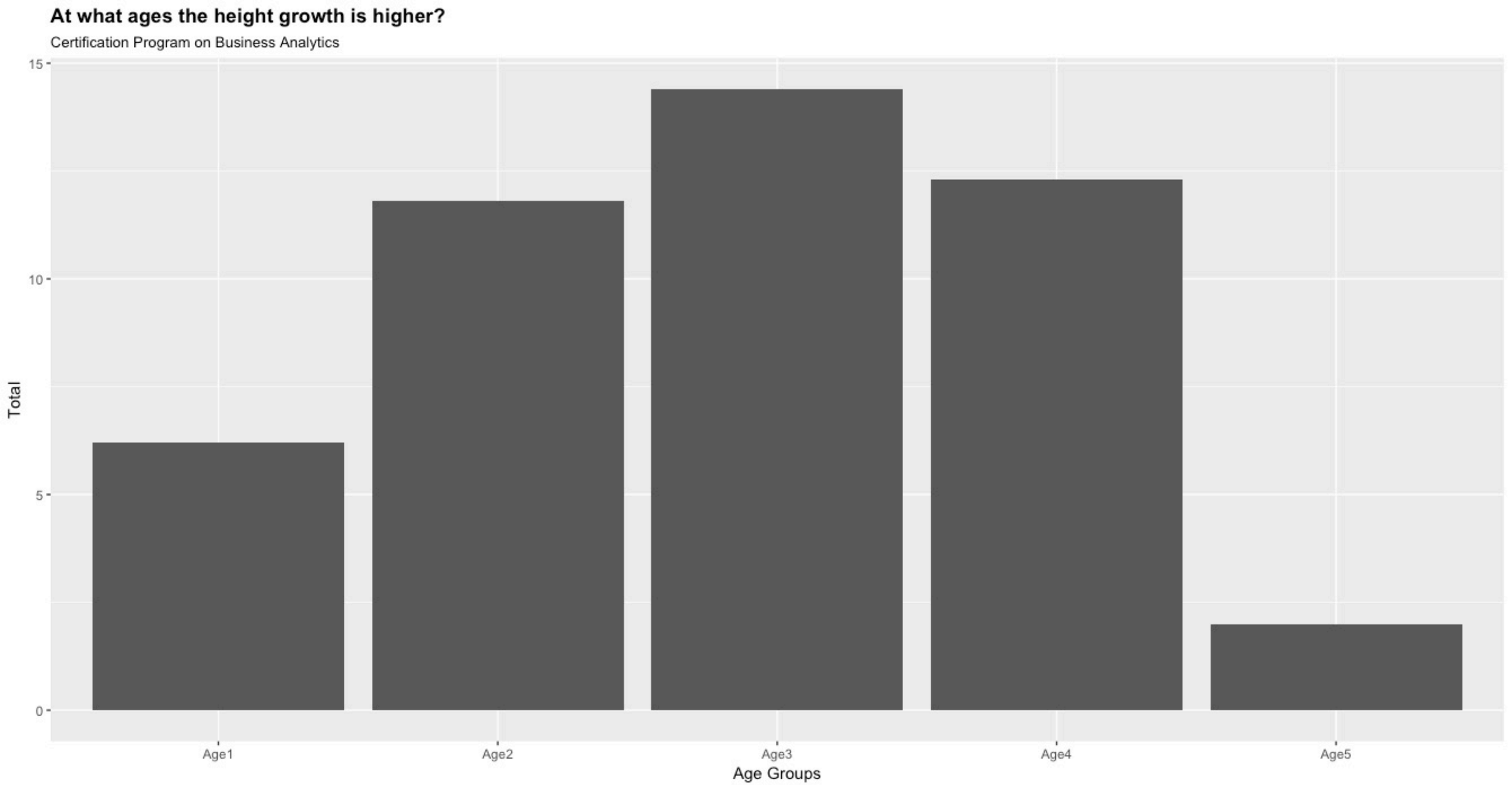
2. Would the height get increased when a person does pyshical activities more in a day?

3. Are the levels of Blood pressures affecting on the height growth in day?. How it would be if a person is under stage 2 prehypertension?

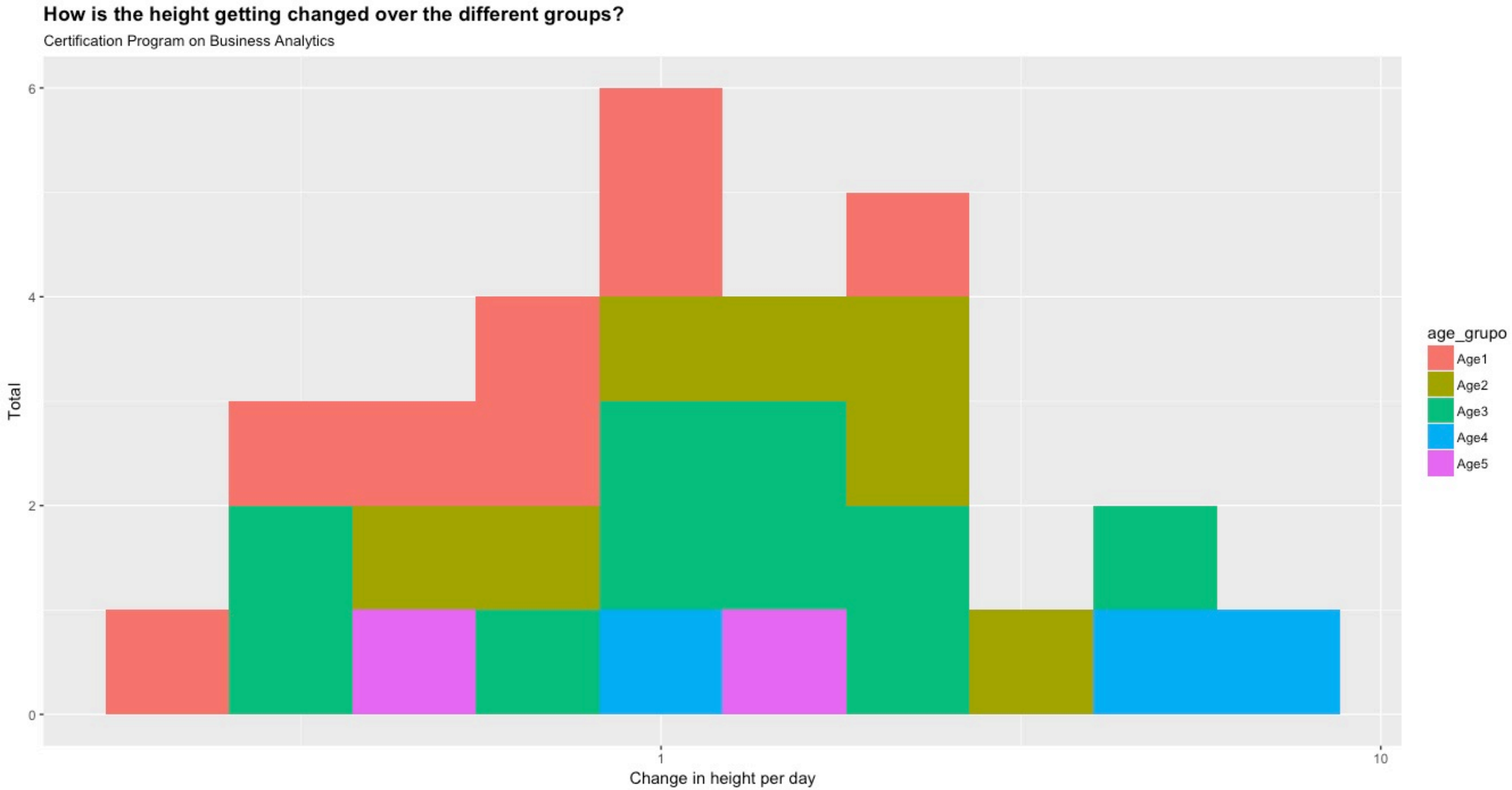
4. Is the height of more happier people's growing higher in a day?

5. How is the height growth in a people who are more obese?

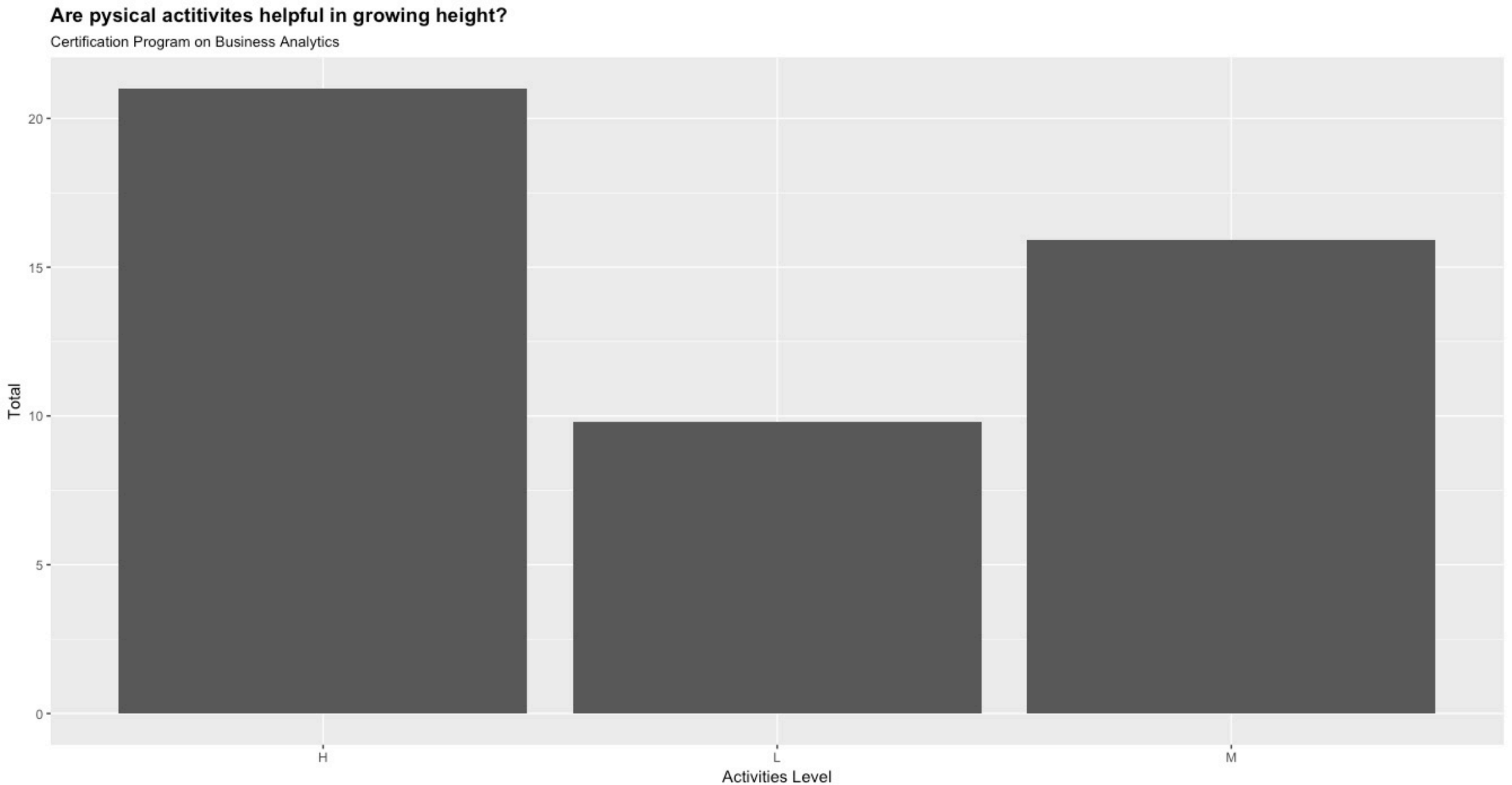
Answer 1: At what ages the height growth is higher?



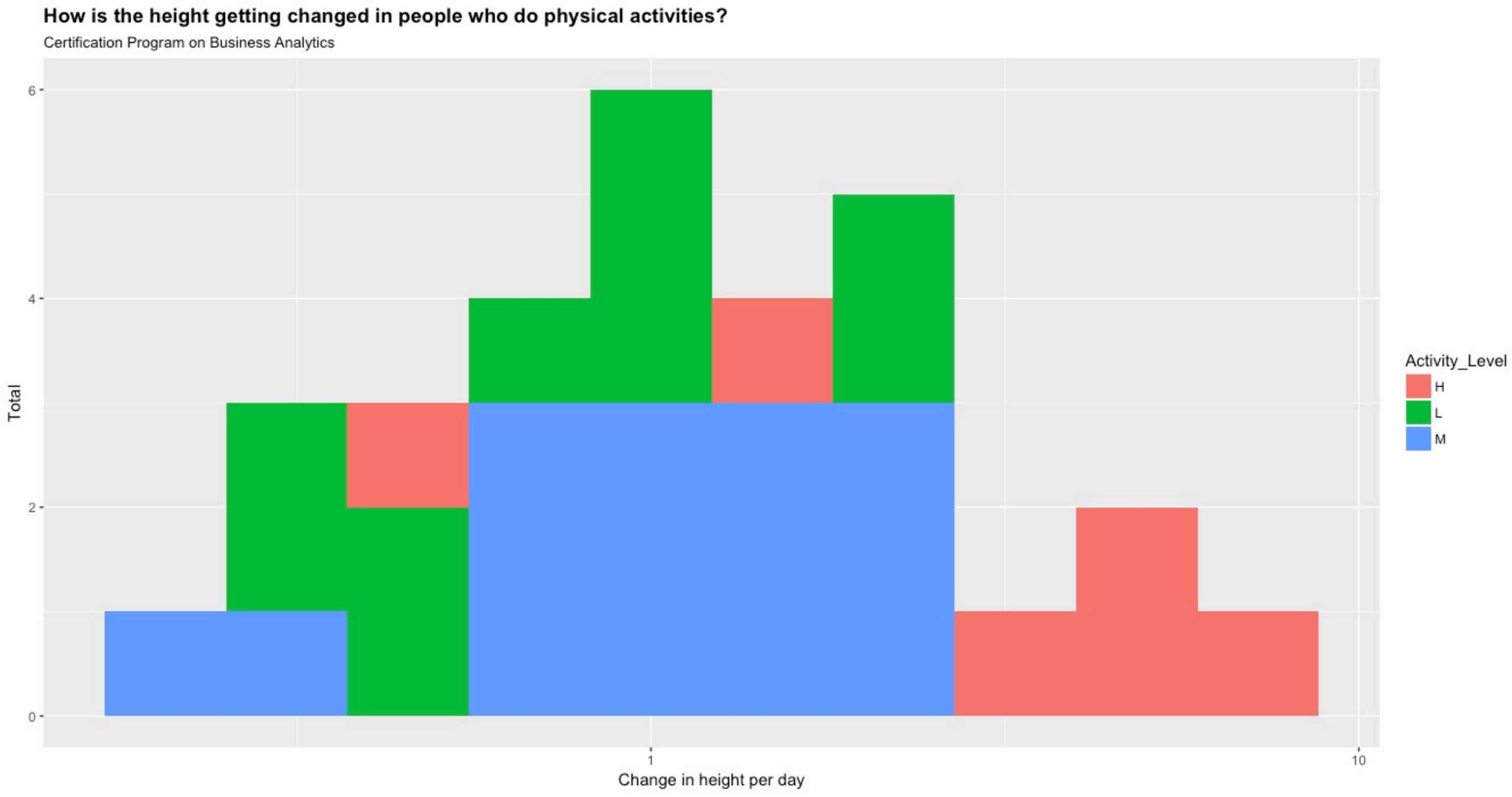
Answer 1: How is the height getting changed over the different groups?



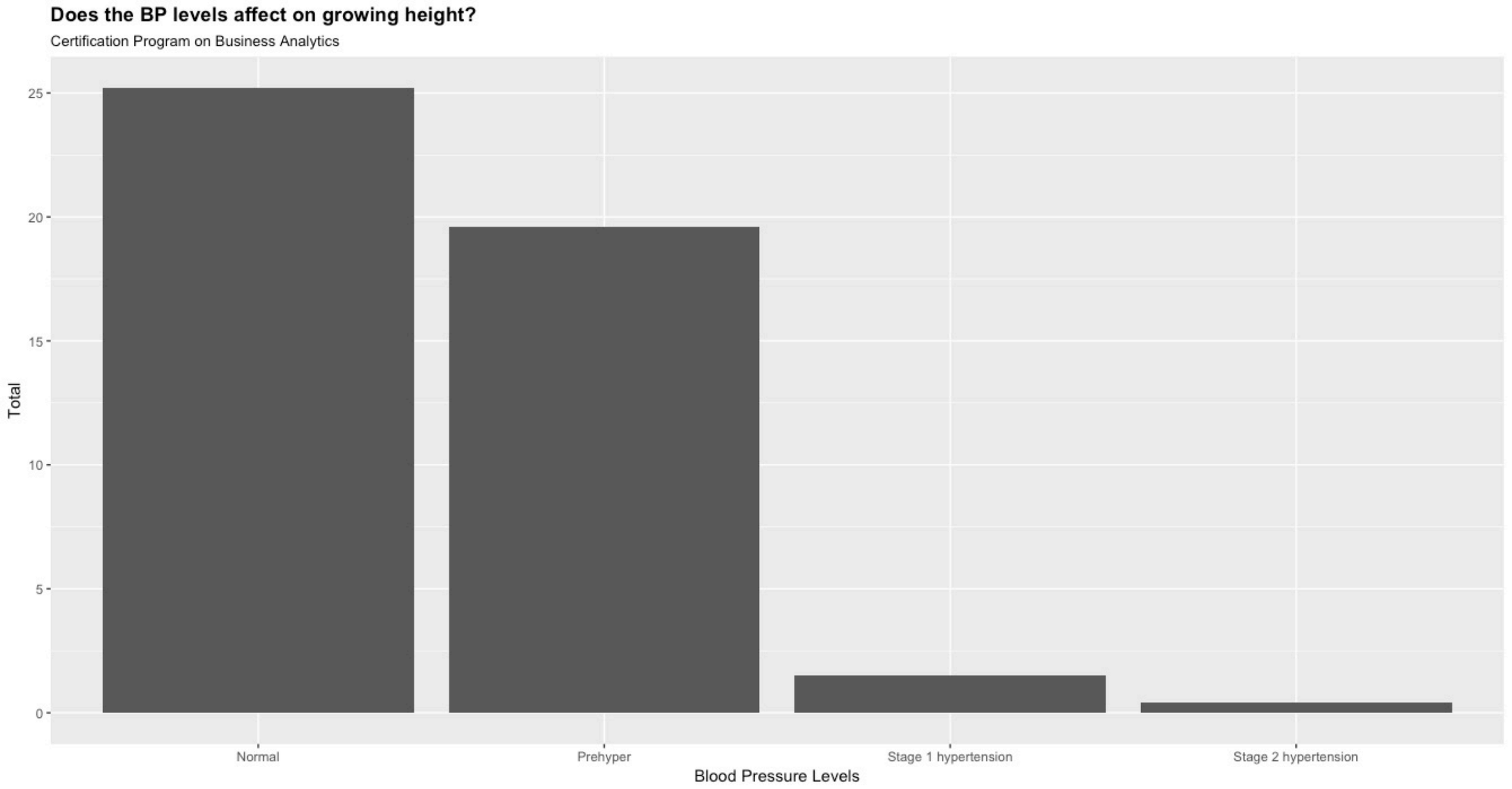
Answer 2: Arey physical activities helpful in growing height?



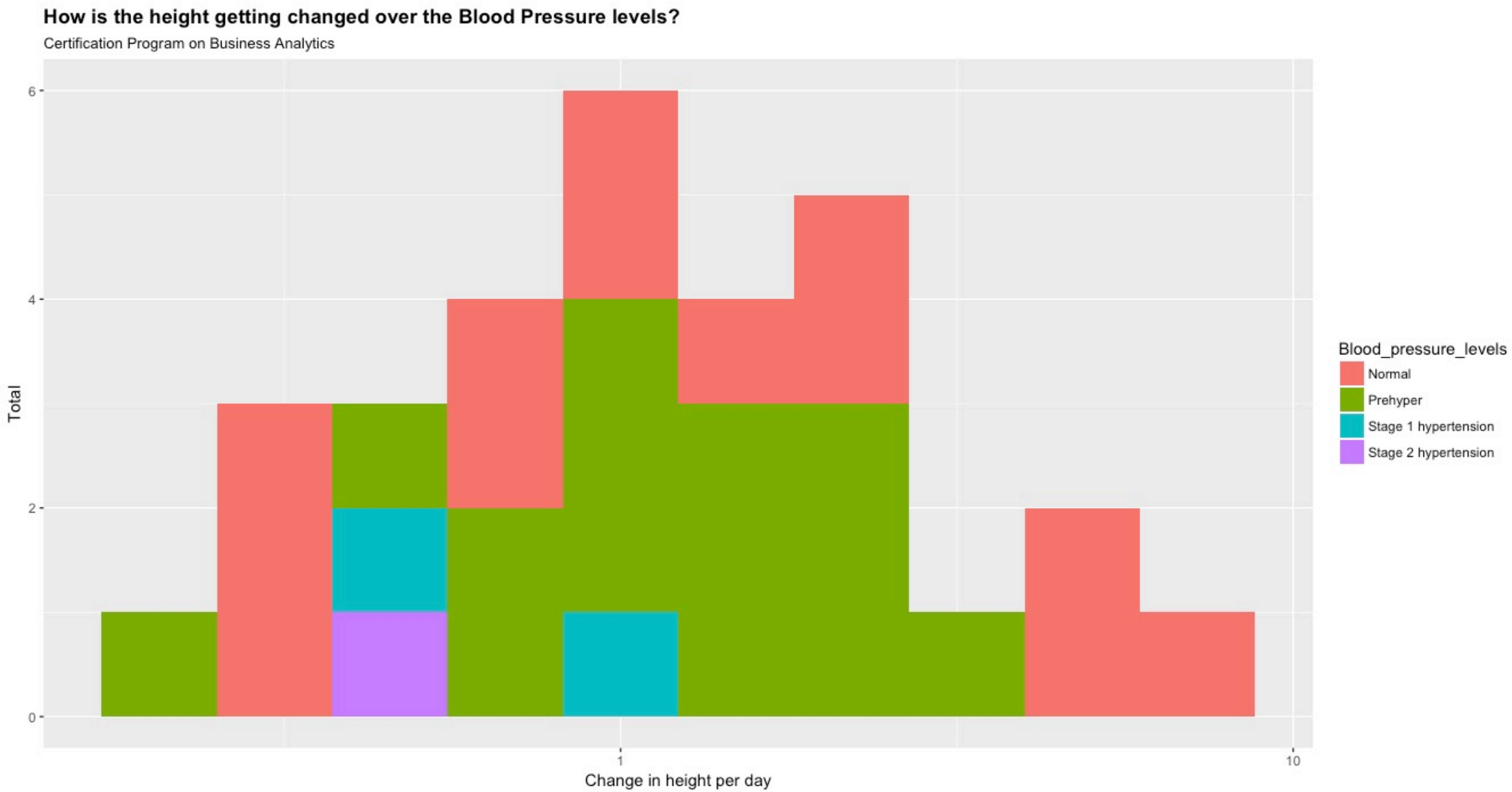
Answer 2: How is the height getting changes in people who do physical activities



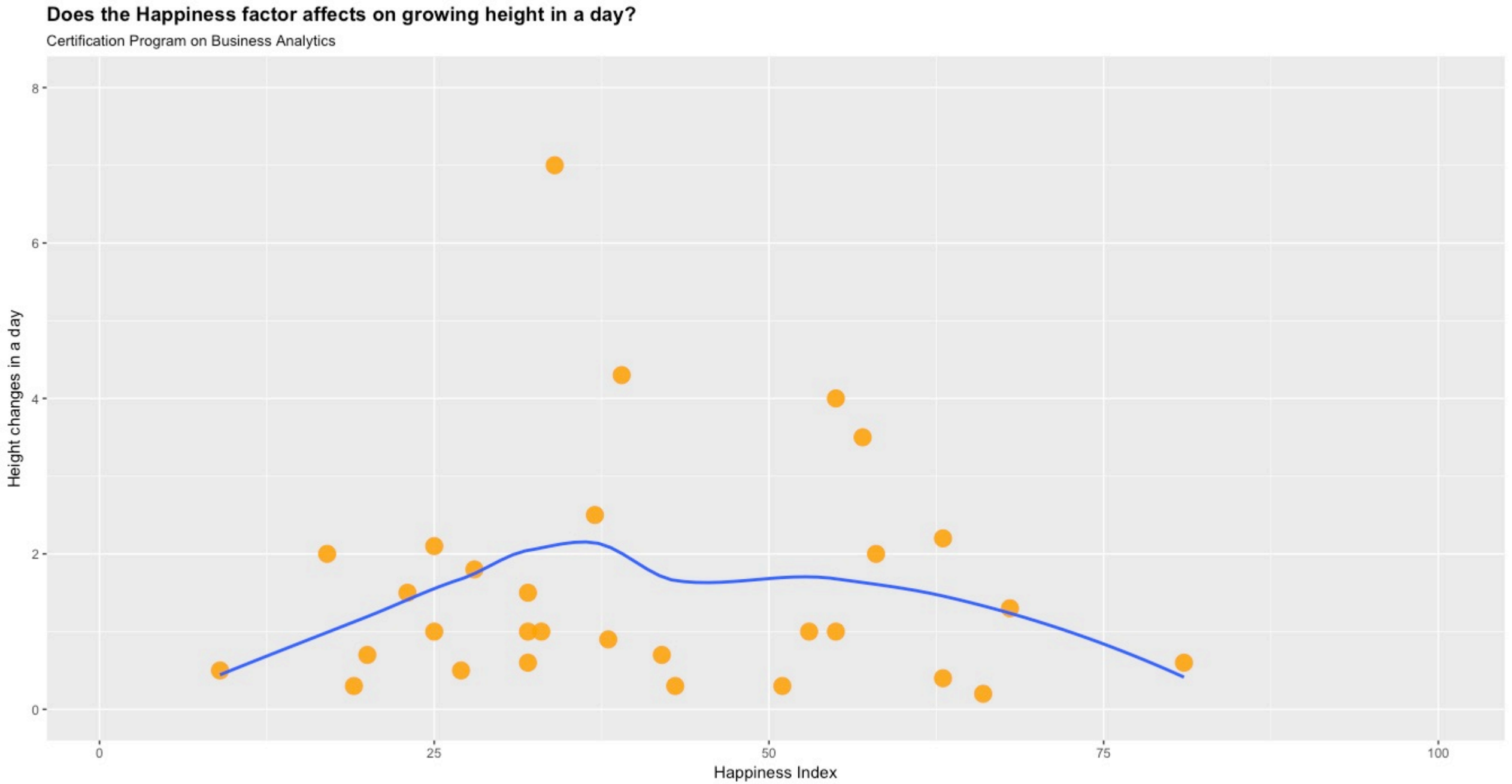
Answer 3: Does the BP levels affect on growing height?



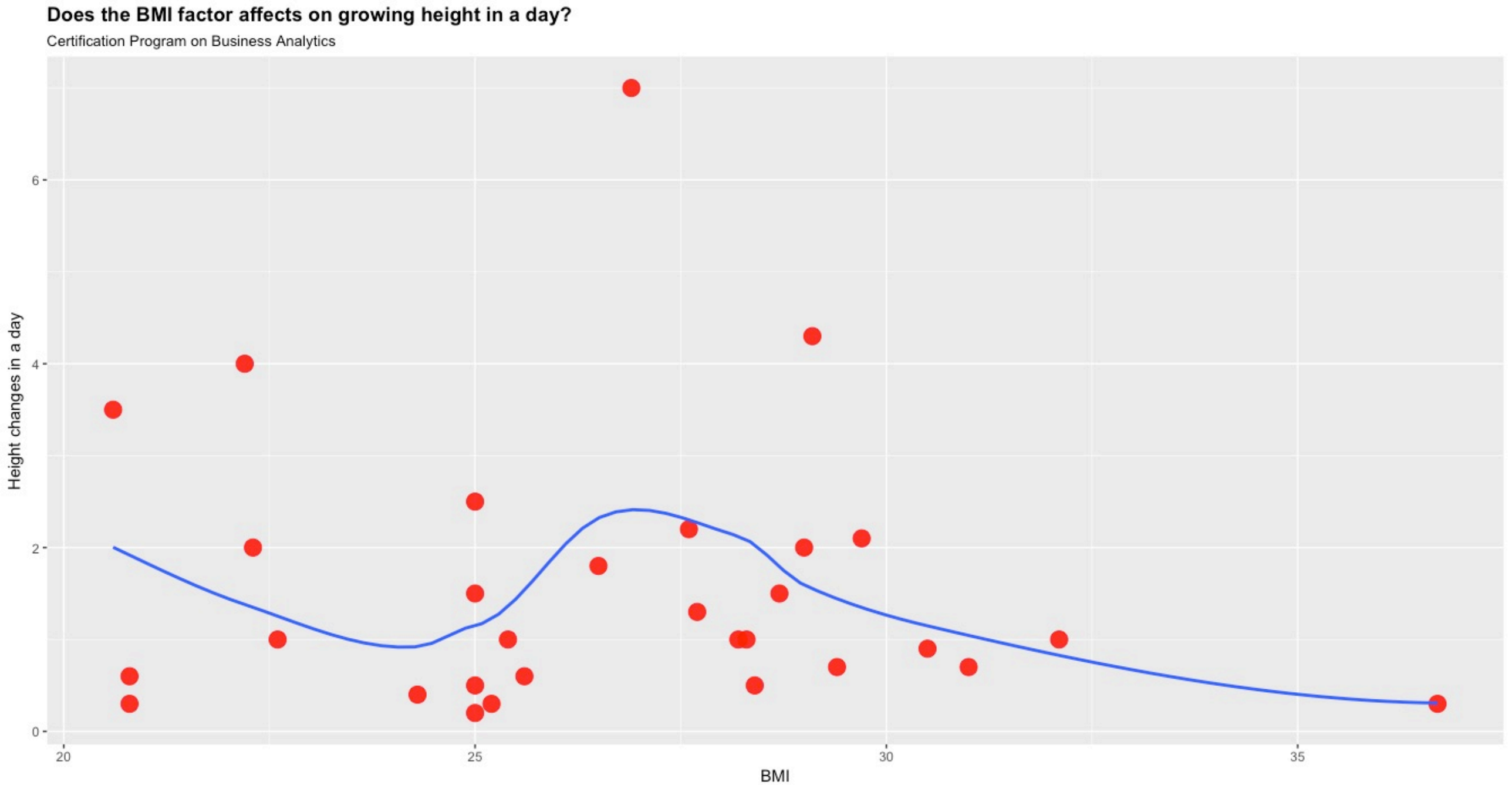
Answer 3: Does the BP levels affect on growing height?



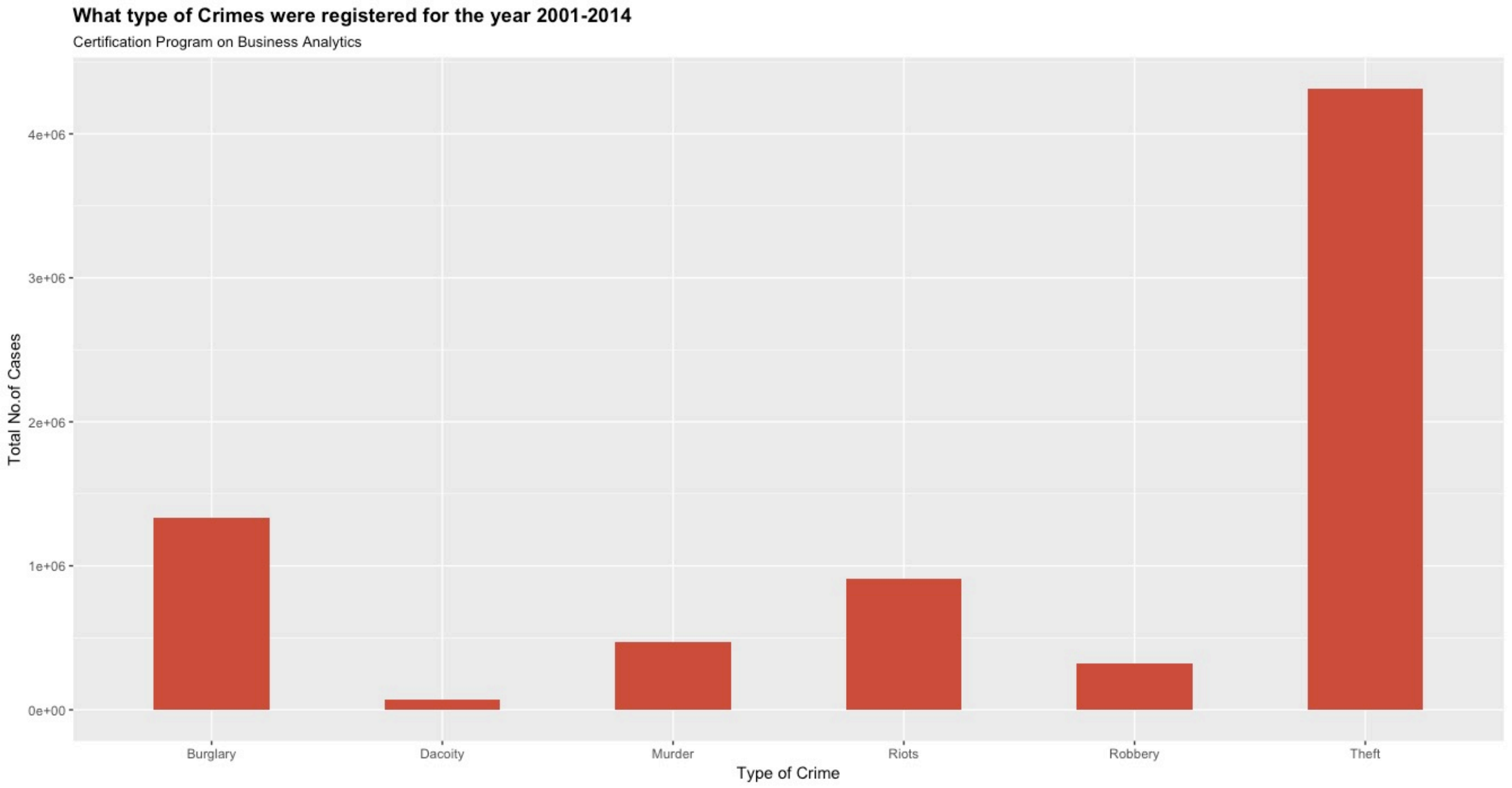
Answer 4: Does the happiness factor affects on growing height in a day?



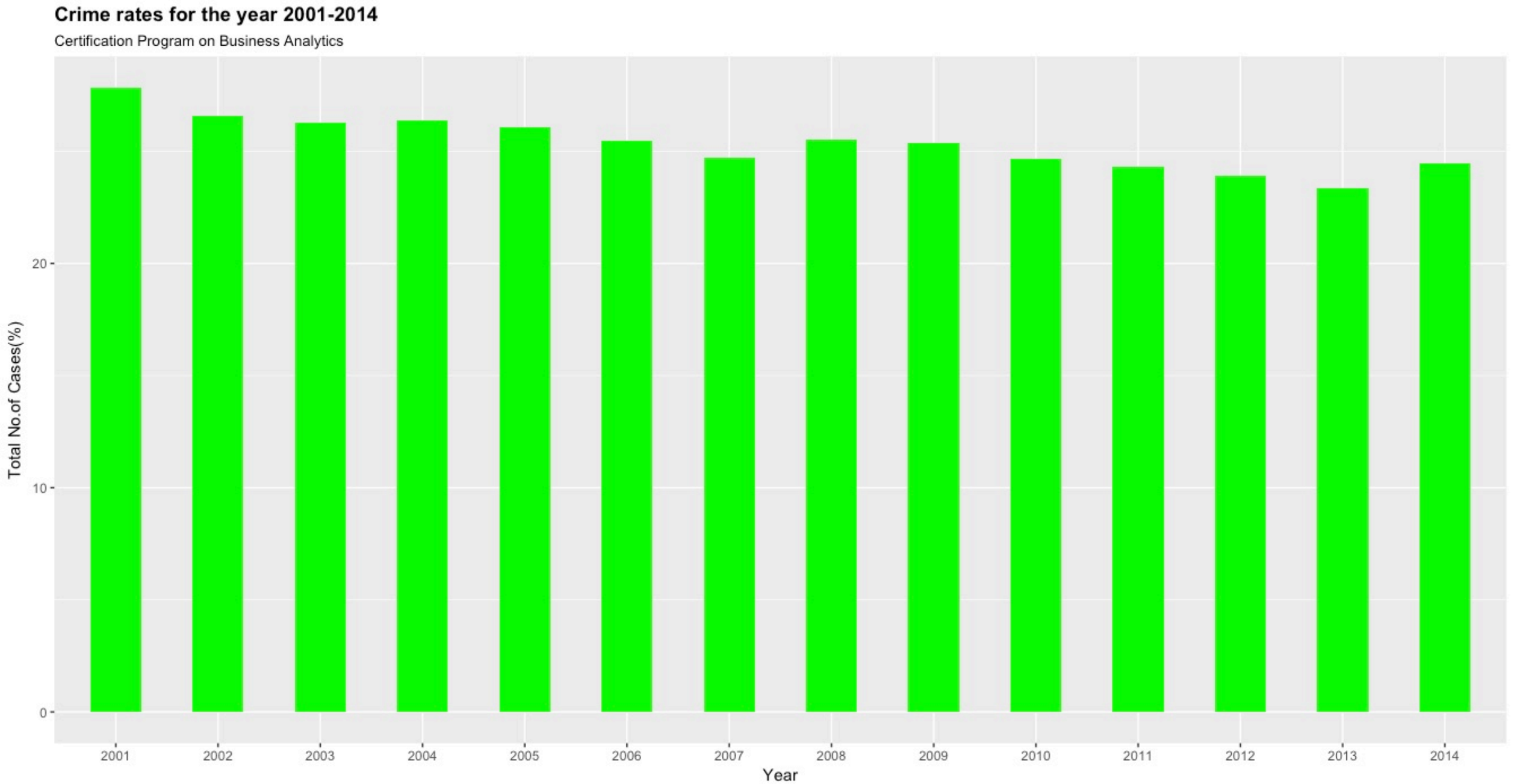
Answer5: Does the BMI factor affects on growing height in a day?



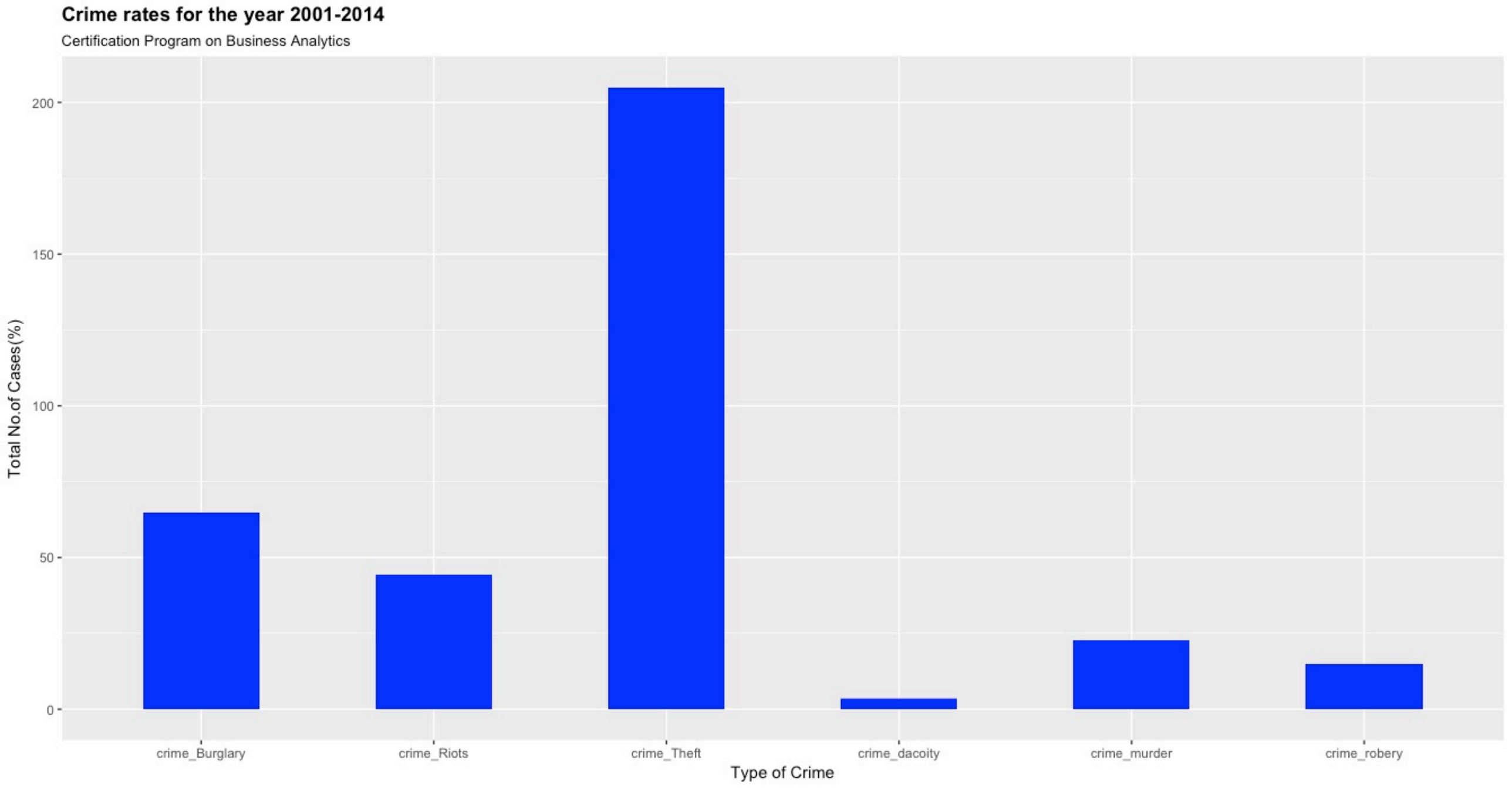
Descriptive analytics on the Indian Crime data



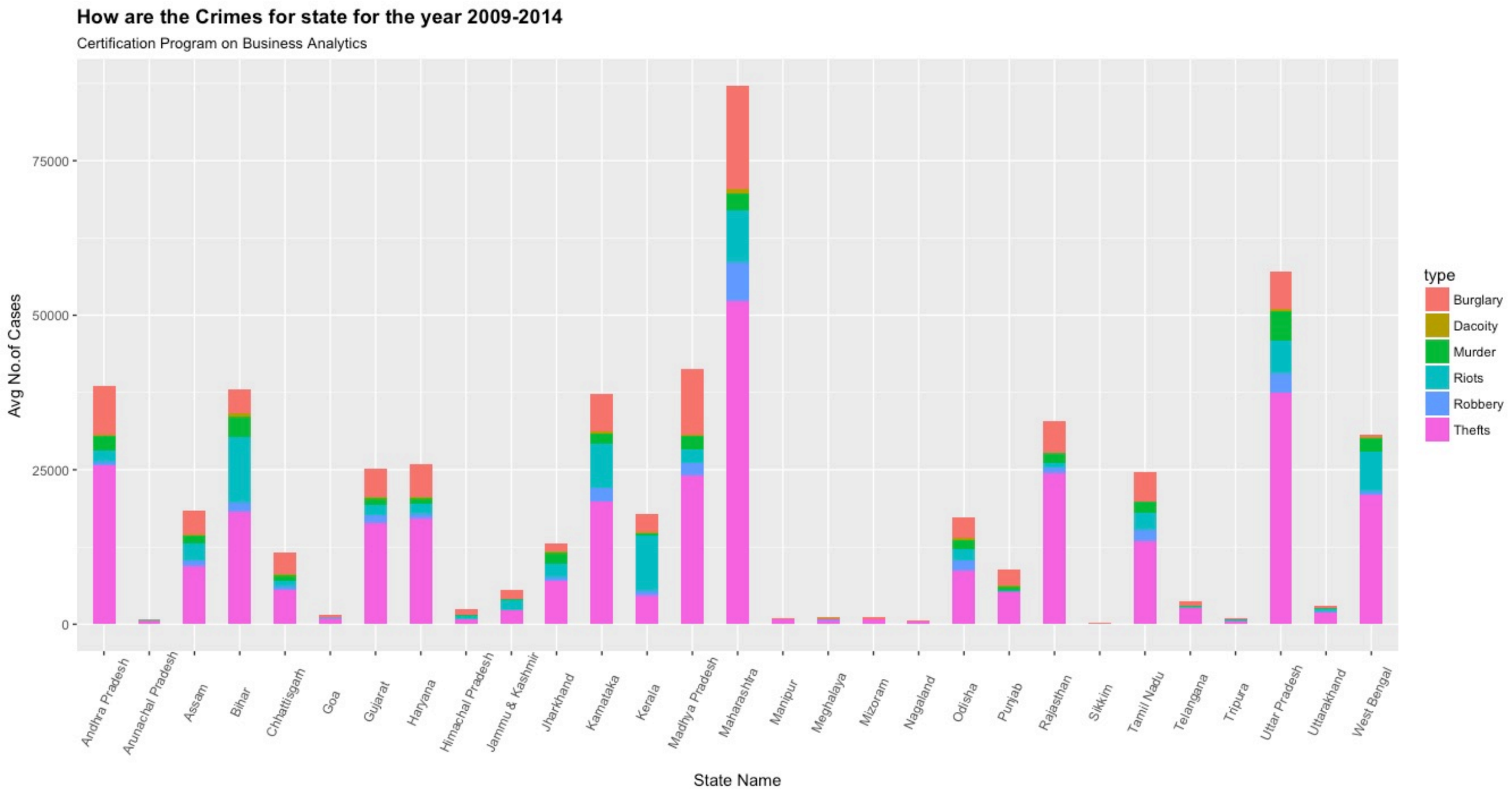
Descriptive analytics on the Indian Crime data



Descriptive analytics on the Indian Crime data

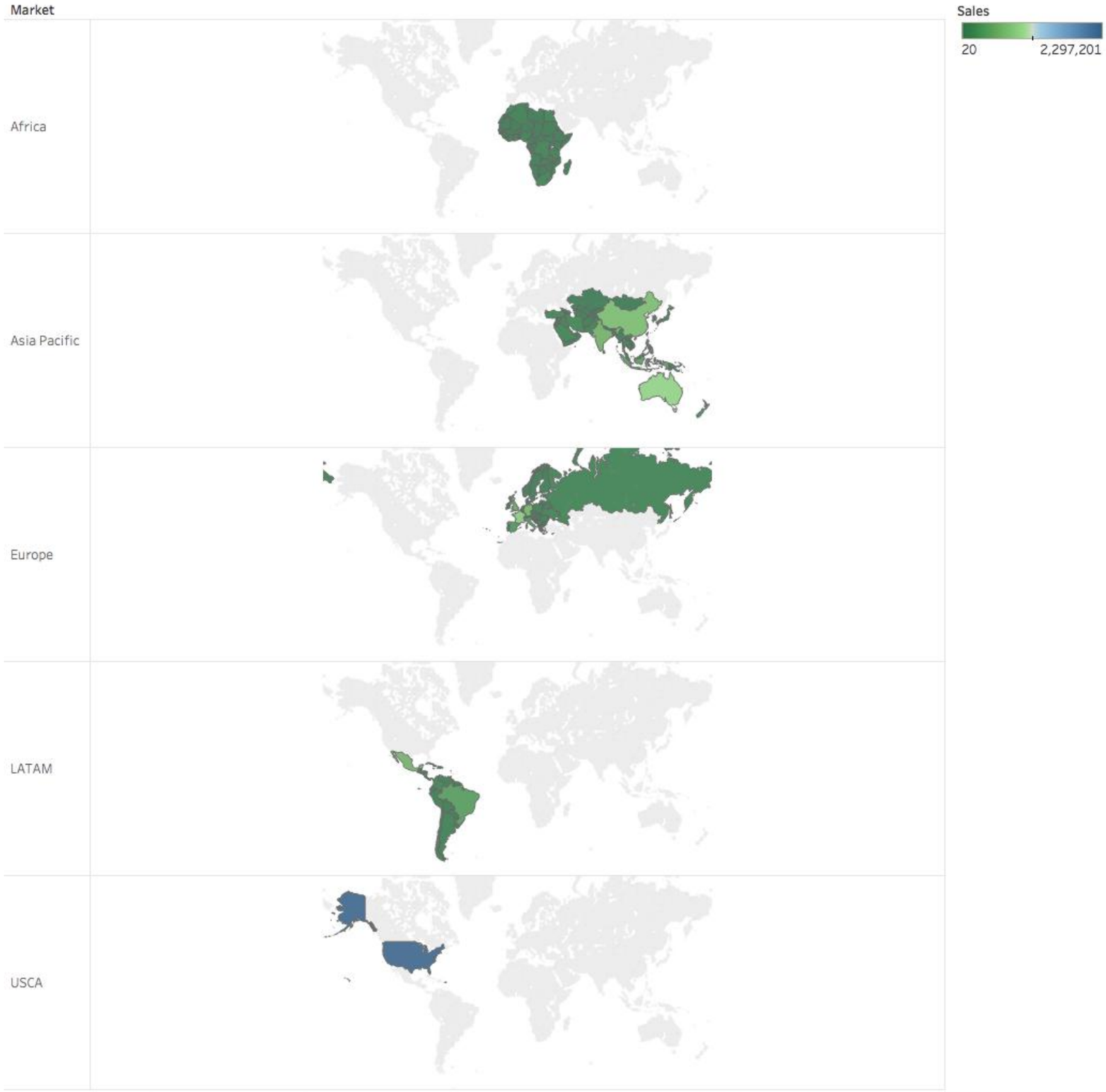


Descriptive analytics on the Indian Crime data



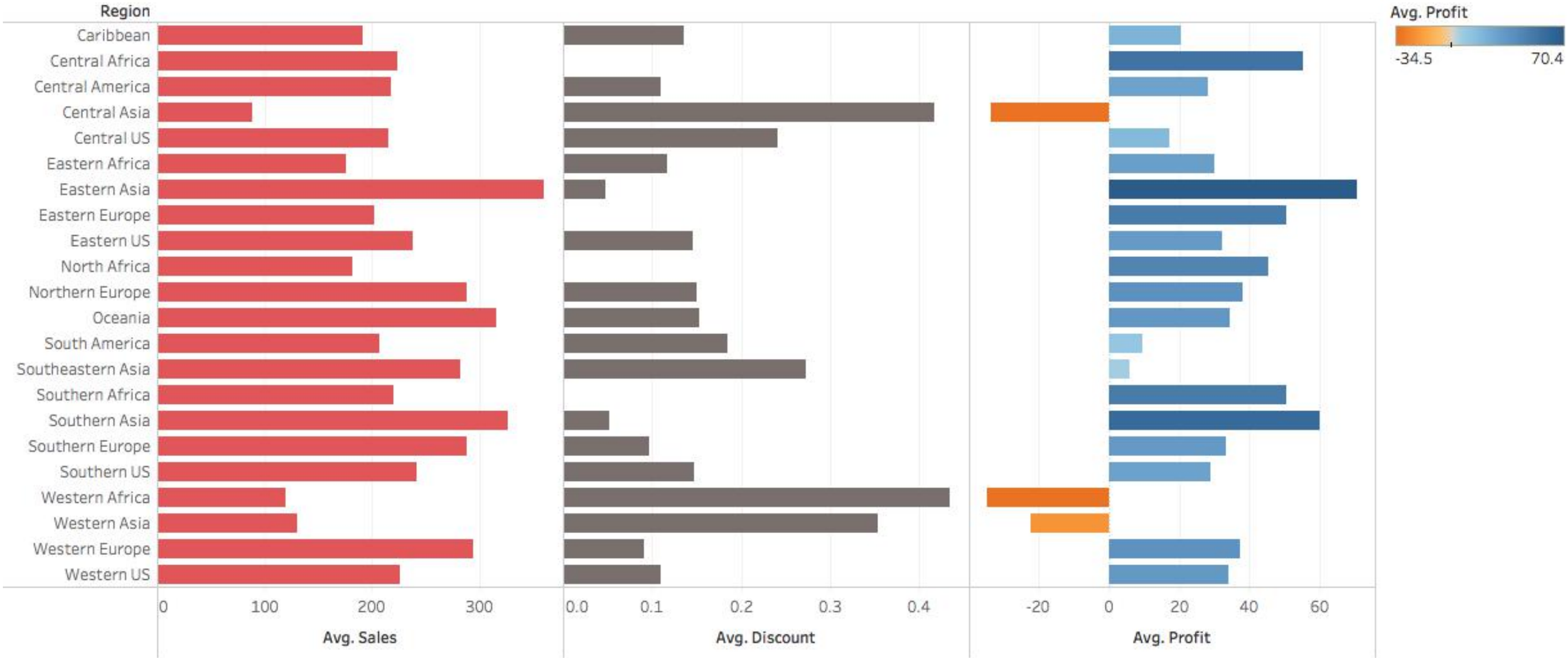
Data Visualiztion Techniques applied on Global Superstore Data Using Tableau

How are the sales per Market World wide?



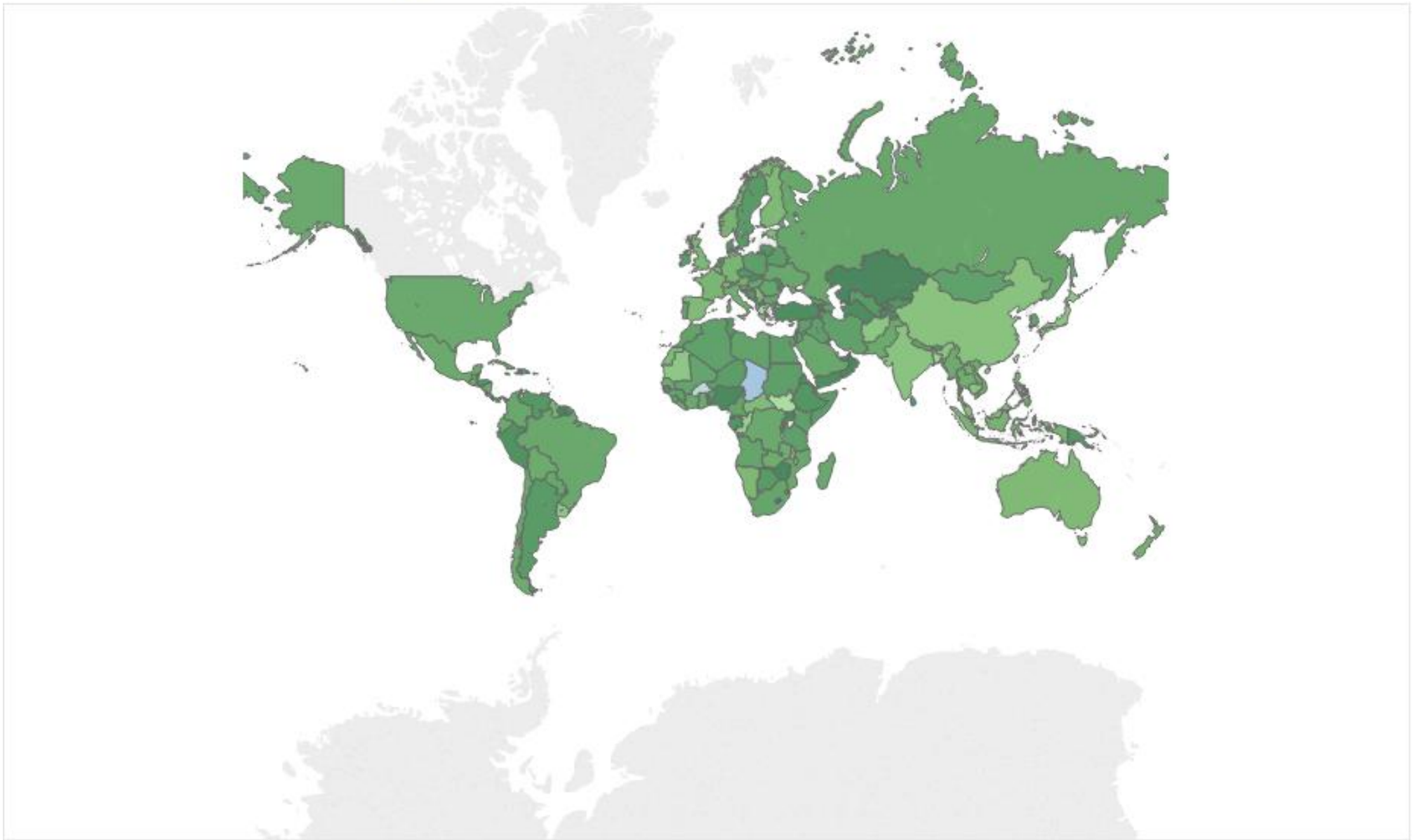
Map based on Longitude (generated) and Latitude (generated) broken down by Market. Color shows sum of Sales. Details are shown for Country.

What are the avg.no of Sales, Discounts and Profits region wise?



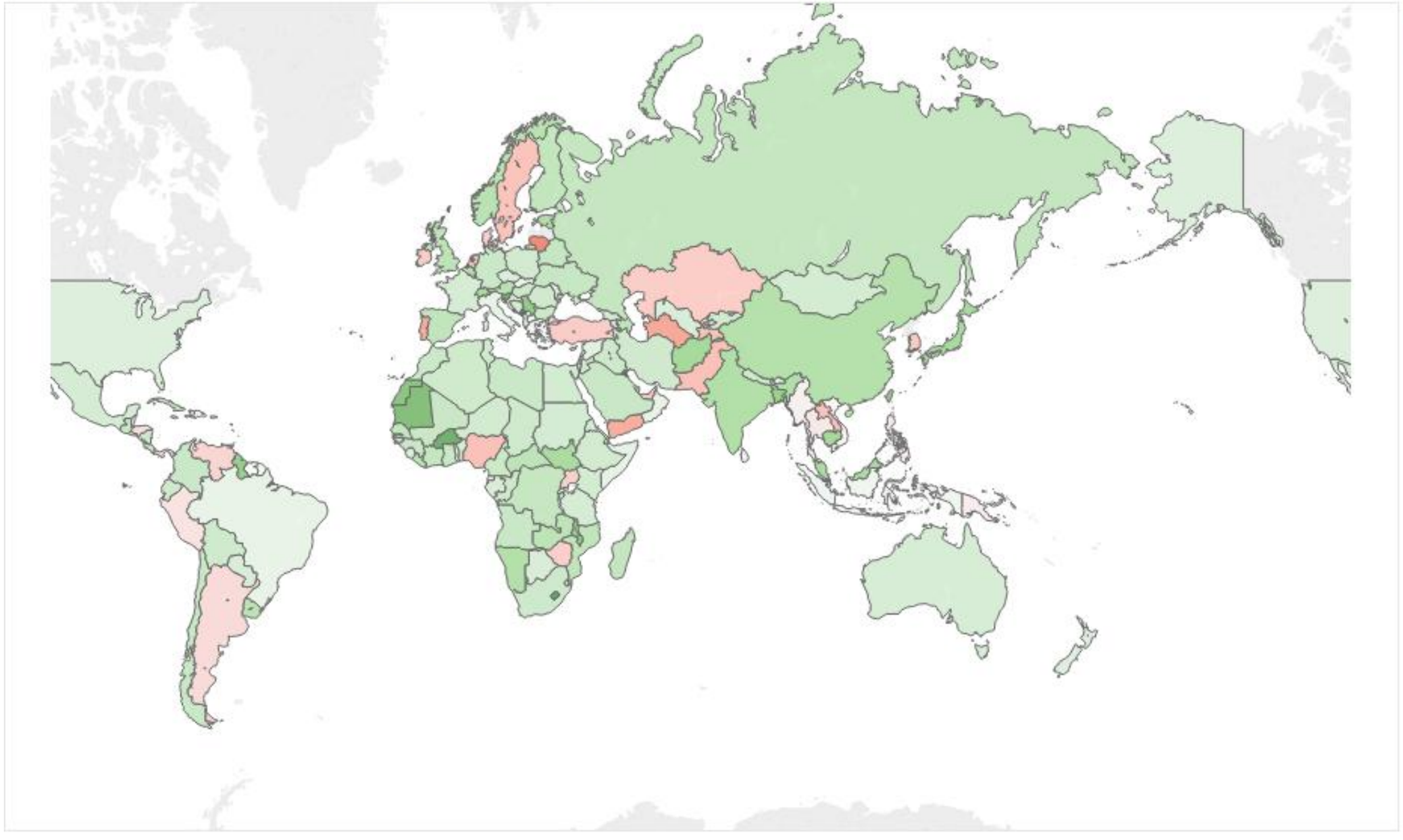
Average of Sales, average of Discount and average of Profit for each Region. For pane Average of Profit: Color shows average of Profit.

How are the avg.no of sales across countries?



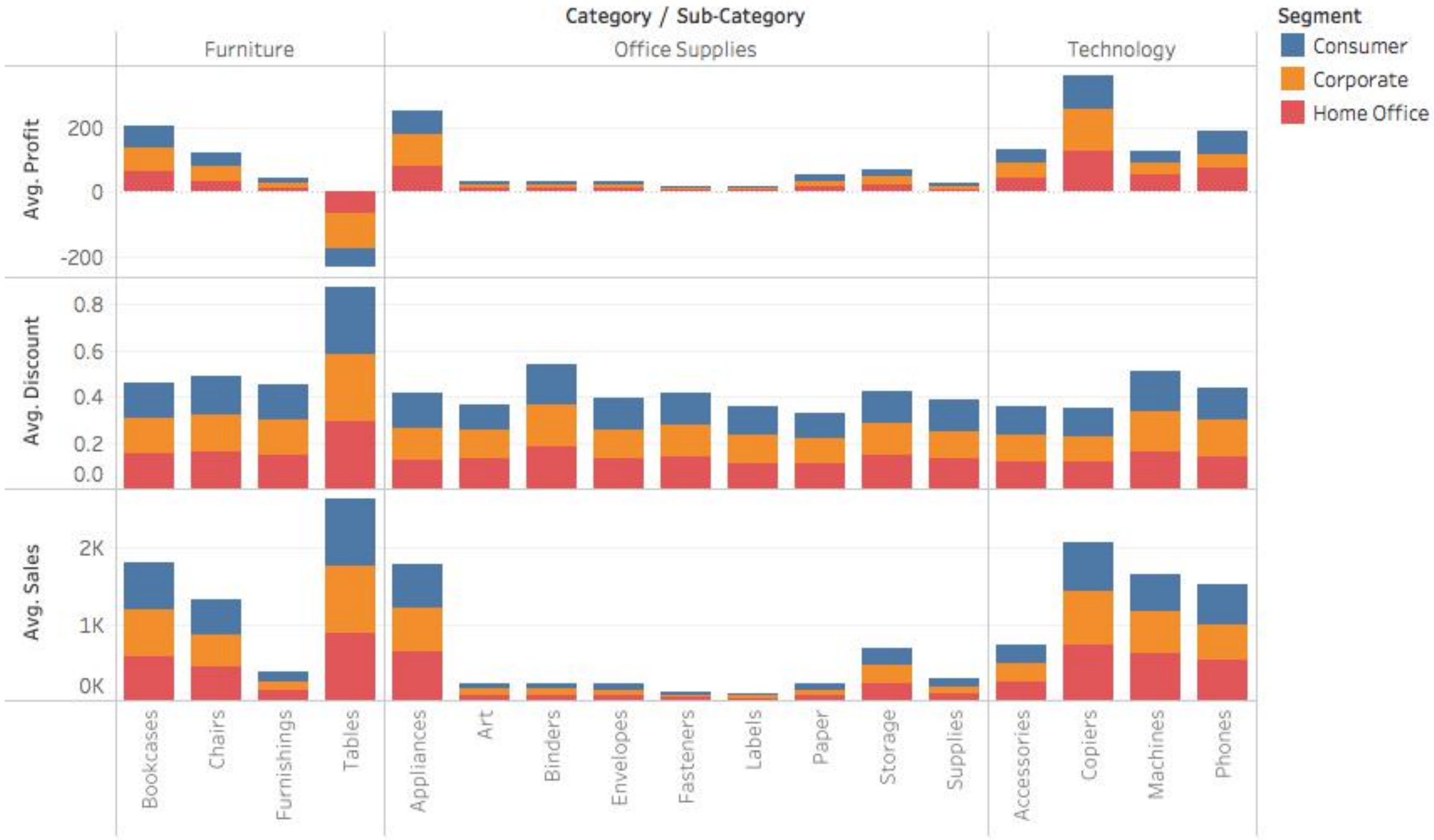
Map based on Longitude (generated) and Latitude (generated). Color shows average of Sales. Details are shown for Country.

How are the avg.no of profits across countries?



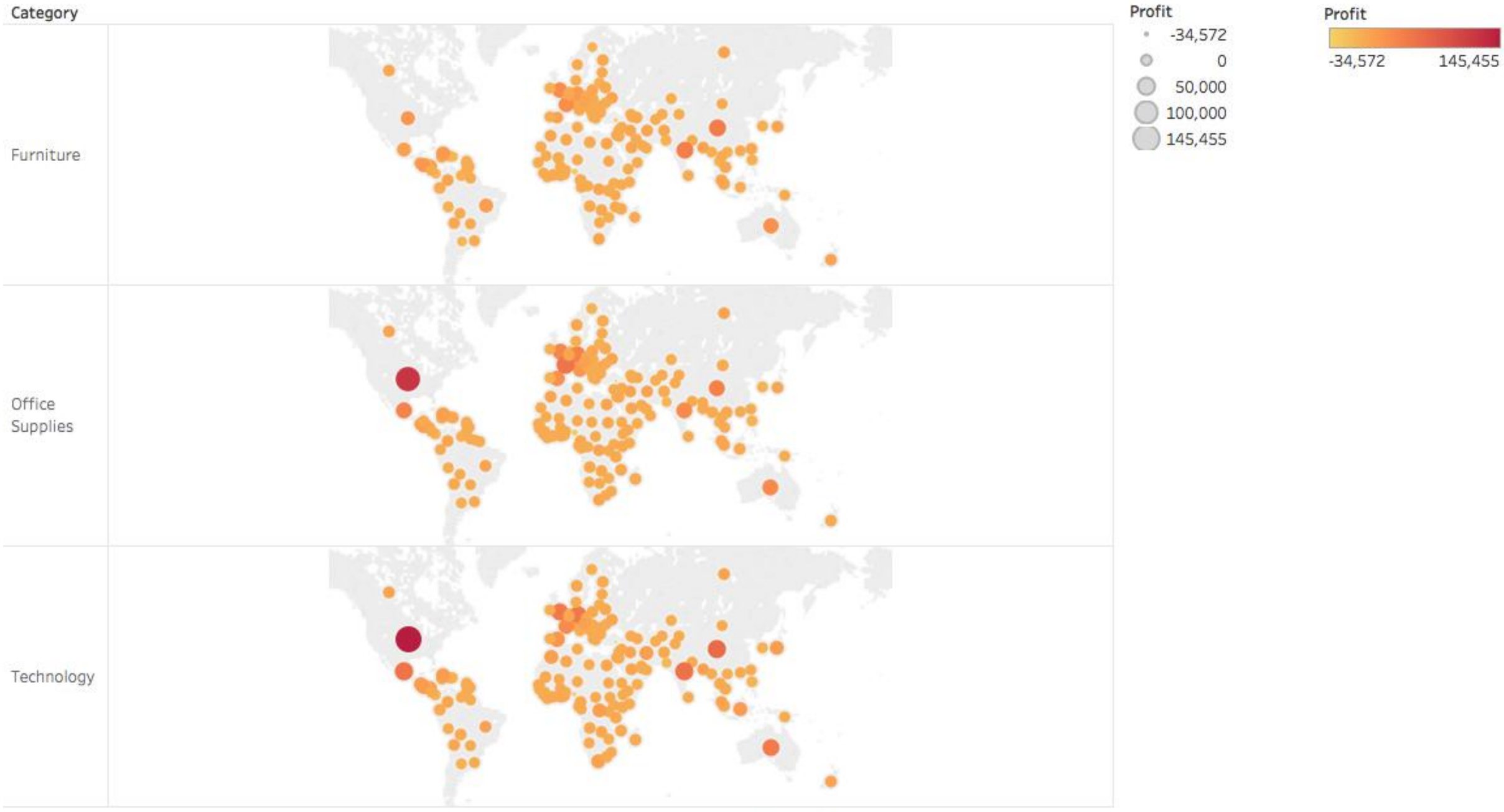
Map based on Longitude (generated) and Latitude (generated). Color shows average of Profit. Details are shown for Country.

Avg No of Sales/Discounts/Profits Per Category in the different segments?



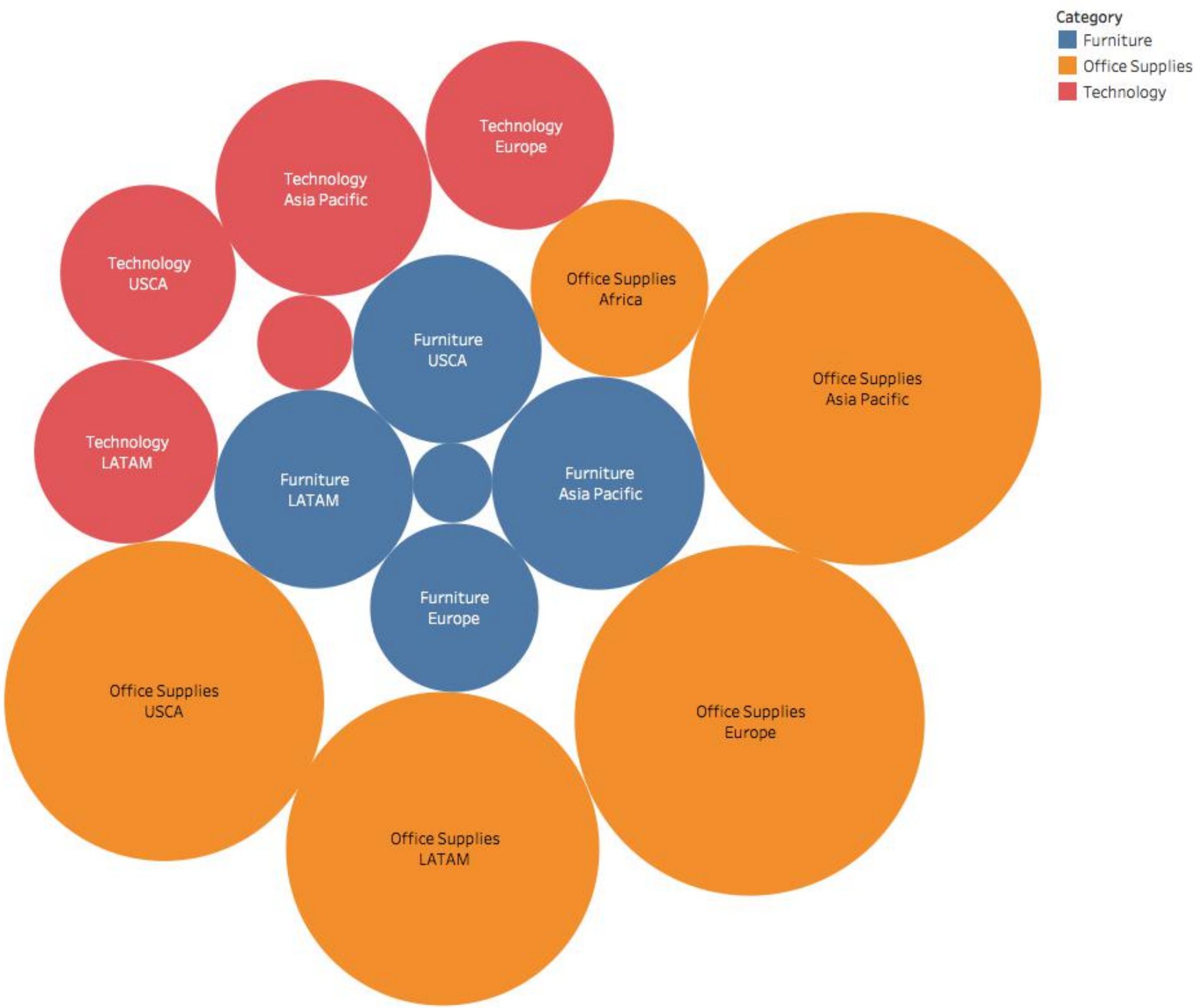
Average of Profit, average of Discount and average of Sales for each Sub-Category broken down by Category. Color shows details about Segment.

Where are the more profits generated per a category?

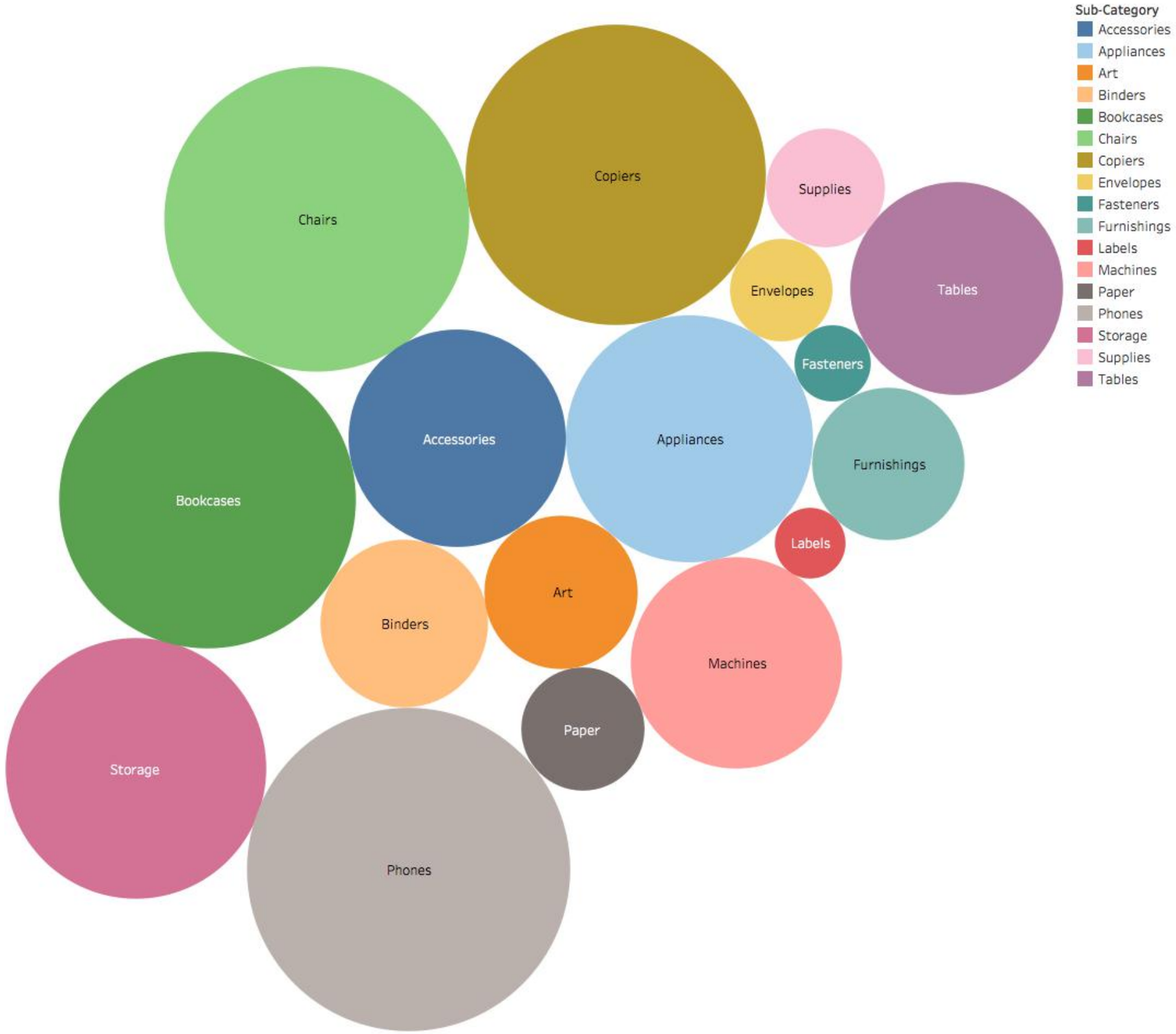


Map based on Longitude (generated) and Latitude (generated) broken down by Category. Color shows sum of Profit. Size shows sum of Profit. Details are shown for Country.

Output of the different category Products per Market Wise??

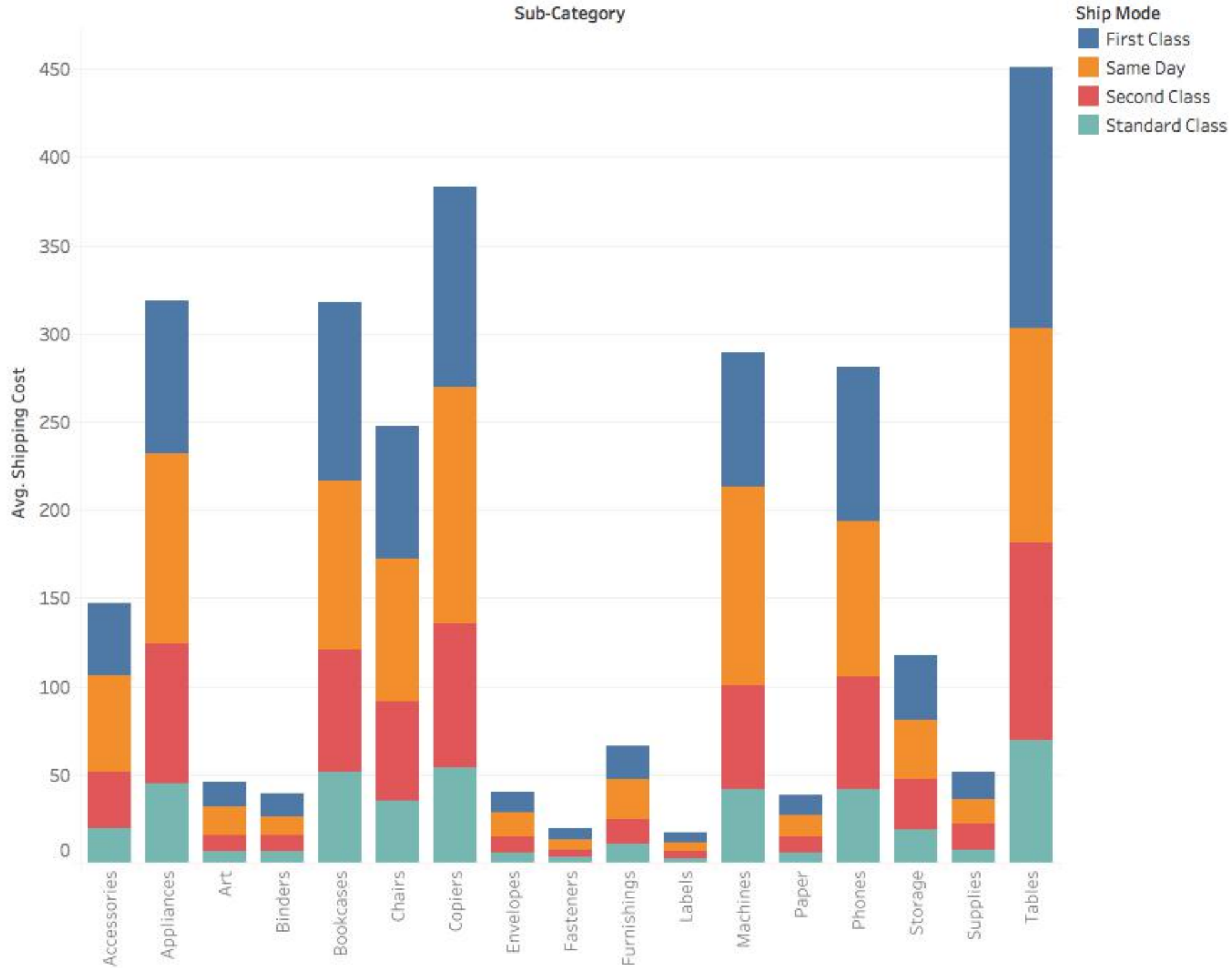


In which category of products we are spending more money for shipping ?



Sub-Category. Color shows details about Sub-Category. Size shows sum of Shipping Cost. The marks are labeled by Sub-Category.

How are the avg. no of shipping cost per a category in each ship mode?



Average of Shipping Cost for each Sub-Category. Color shows details about Ship Mode.

Thanks you (Muchas gracias)!!!