Iván Daniel Hernández Rocha

id.hernandezrocha@ugto.mx

Abstract

Análisis del archivo survey\_results.csv, que contiene la información de una encuesta a los usuarios de Stack Overflow en 2019. La información concierne las variables de: país, nivel de educación, tipo de desarrollador, años de experiencia, salario anual (US dollars), horas de trabajo a la semana, lenguajes de programación, edad, género y etnia.

minería de datos

Práctica

1. Al realizar un análisis preliminar (five-number summary, boxplot) de la variable para el salario anual (**ConvertedComp**) con respecto a los tres géneros disponibles dentro de la variable **Gender** (Man, Woman, Non-binary, genderqueer, or gender non-conforming) podemos observar los siguientes resultados:

Para la variable **Gender = Man**:

Chart, box and whisker chart

Description automatically generated

Graphical user interface, text, chat or text message

Description automatically generated

Para la variable **Gender = Woman**:

Chart, box and whisker chart

Description automatically generated

Text

Description automatically generated

Para la variable **Gender = Non-binary, genderqueer, or gender non-conforming**:

Chart, box and whisker chart

Description automatically generated

Text

Description automatically generated

Haciendo un recorte de 10% a los datos, podemos observar cambios muy ligeros dentro del boxplot, lo cual significa que no tenemos valores extremos que puedan interferir en la correcta interpretación de los datos.

A continuación, se muestran los datos originales y recortados al 10% para su comparación:

**Gender = Man**

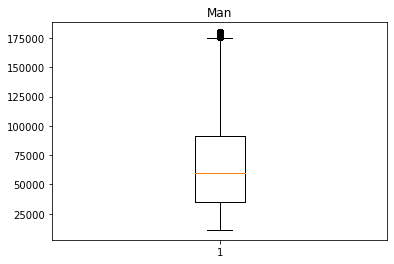
Graphical user interface, text, chat or text message

Description automatically generatedText

Description automatically generated

Datos originales Datos recortados 10%

Chart, box and whisker chart

Description automatically generated

Datos originales Datos recortados 10%

**Gender = Woman**

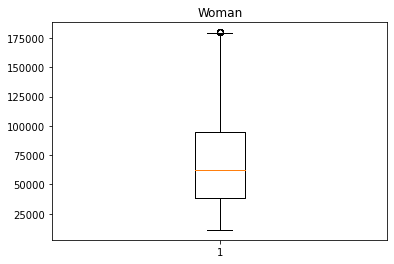
Text

Description automatically generatedText

Description automatically generated

Datos originales Datos recortados 10%

Chart, box and whisker chart

Description automatically generated

Datos originales Datos recortados 10%

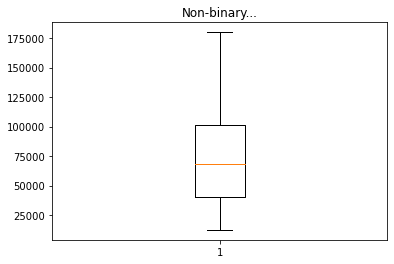
**Gender =** **Non-binary, genderqueer, or gender non-conforming**

Text

Description automatically generatedText

Description automatically generated

Datos originales Datos recortados 10%



Chart, box and whisker chart

Description automatically generated

Datos originales Datos recortados 10%

El cambio más significativo con respecto a los datos originales se ve en el genéro de Non-binary, genderqueer, or gender non-conforming, donde la media y la mediana aumentan en aproximadamente $2,000.

1. En la siguiente gráfica podemos observar que el género con mas respuestas es ‘Man’ con 27,062.

Chart, pie chart

Description automatically generated

1. Haciendo un análisis de los salarios anuales promedio, podemos observar que el género que tiende a tener los salarios más altos es ‘Non-binary, genderqueer, or gender non-conforming’, con un salario promedio de $76,178.59 y a su vez, el género que tiende a tener los salarios más bajos es ‘Man’ con un salario promedio de $66,791.98.

Chart, bar chart

Description automatically generated

1. El análisis de los datos también nos muestra que los lenguajes de programación más populares y menos populares para cada género son los que se muestran a continuación:

**Gender = Man**

Chart, pie chart

Description automatically generated

**El lenguaje más popular es: JavaScript.**

**El lenguaje menos popular es Web Assembly**

**Gender = Woman**

Chart, pie chart

Description automatically generated

**El lenguaje más popular es: JavaScript**

**El lenguaje menos popular es: Erlang**

**Gender = Non-binary, genderqueer, or gender non-conforming**

Chart, pie chart

Description automatically generated

**El lenguaje más popular es: JavaScript**

**El lenguaje menos popular es: F#**

1. Al analizar los datos de la variable **DevType,** podemos encontrar cuales son los tipos de desarrollador más comunes para cada género:

**Gender = Man**

Chart, pie chart

Description automatically generated

**El tipo de desarrollador más popular es: Developer, full-stack.**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

**Gender = Woman**

Chart, pie chart

Description automatically generated

**El tipo de desarrollador más popular es: Developer, full-stack**

**El tipo de desarrollador menos popular es: Senior executive/VP**

**Gender = Non-binary, genderqueer, or gender non-conforming**

Chart, pie chart

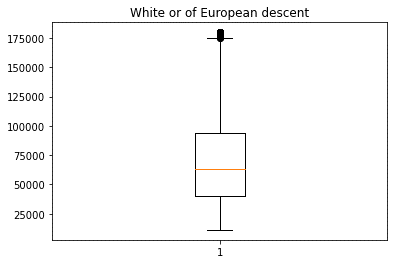
Description automatically generated

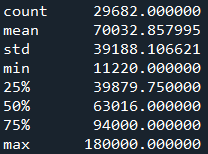
**El tipo de desarrollador más popular es: Developer, full-stack**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

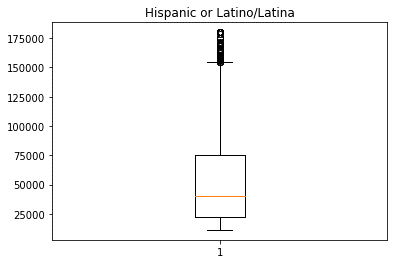
1. Al realizar un análisis preliminar (five-number summary, boxplot) de la variable para el salario anual (**ConvertedComp**) con respecto a las 4 etnias más populares podemos observar los siguientes resultados:

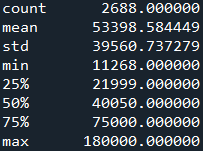
Para la variable **Ethnicity = White or of European descent**:



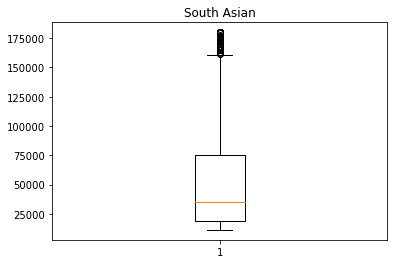


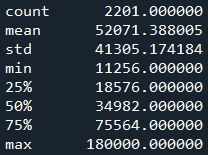
Para la variable **Ethnicity = Hispanic or Latino/Latina**:



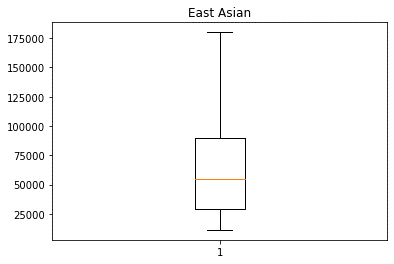


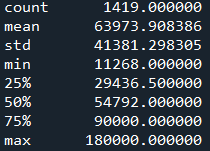
Para la variable **Ethnicity = South Asian**:





Para la variable **Ethnicity =** **East Asian**:

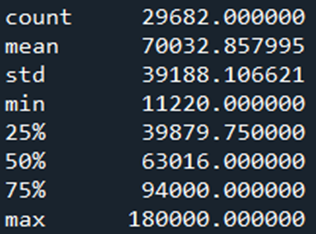
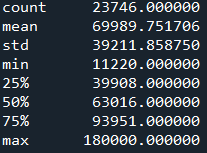




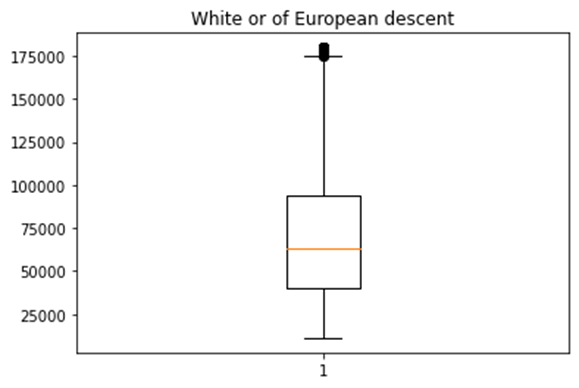
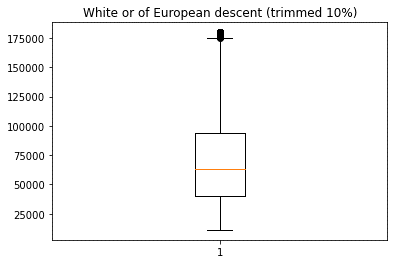
Haciendo un recorte de 10% a los datos, podemos observar un ligero cambio dentro del boxplot para ‘Hispanic or Latino/Latina’ y ‘South Asian’, lo cual significa que se eliminan algunos valores extremos que podrían interferir en la correcta interpretación de los datos.

A continuación, se muestran los datos originales y recortados al 10% para su comparación:

**Ethnicity = White or of European descent**

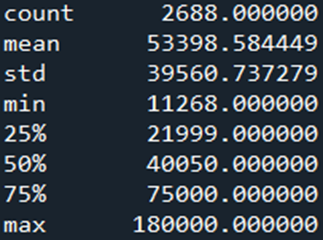
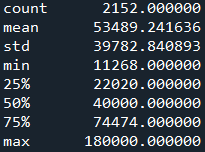


Datos originales Datos recortados 10%

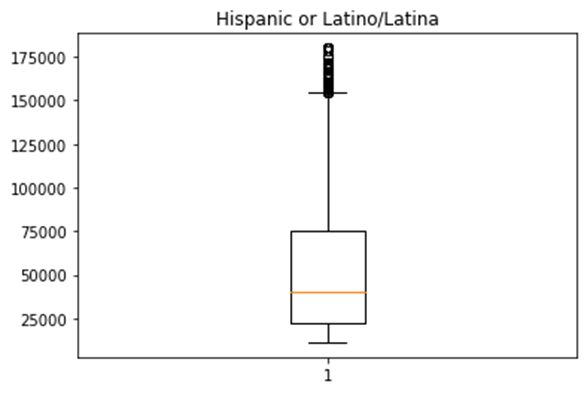
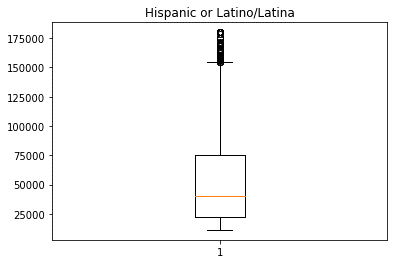


Datos originales Datos recortados 10%

**Ethnicity = Hispanic or Latino/Latina**

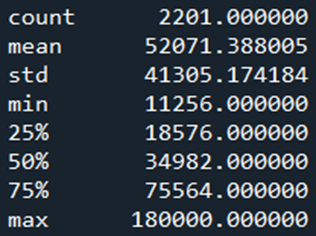
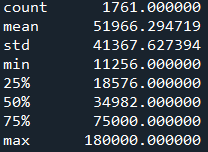


Datos originales Datos recortados 10%

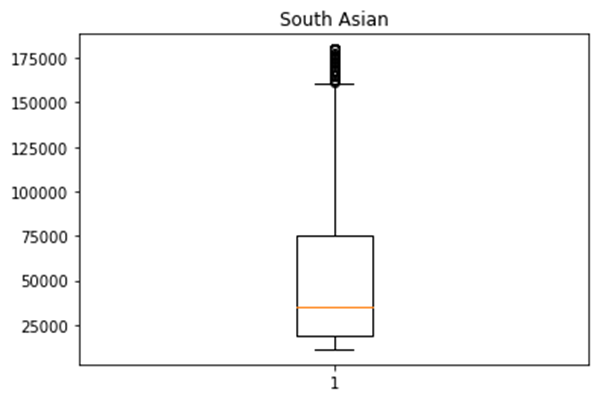


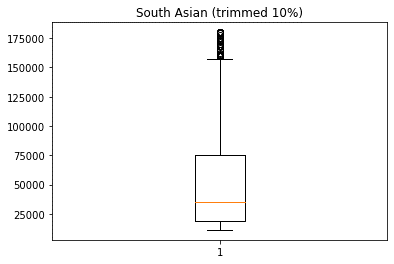
Datos originales Datos recortados 10%

**Ethnicity = South Asian**



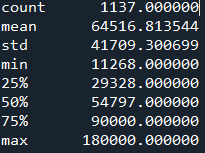
Datos originales Datos recortados 10%

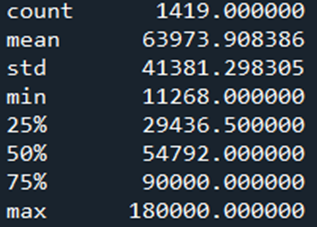




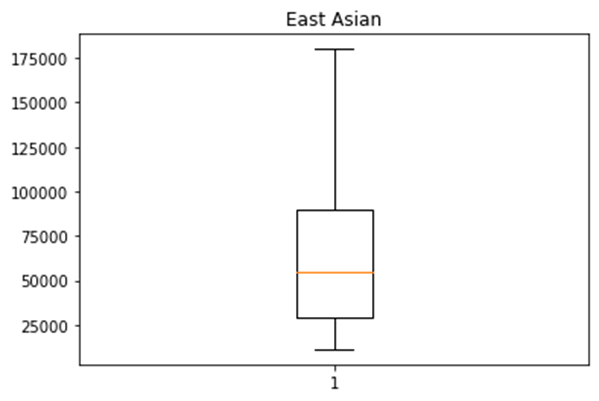
Datos originales Datos recortados 10%

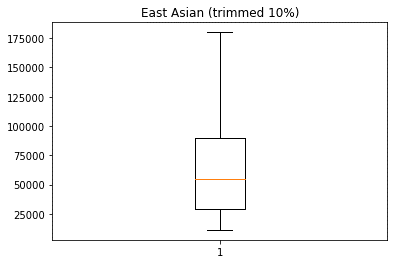
**Ethnicity = East Asian**





Datos originales Datos recortados 10%

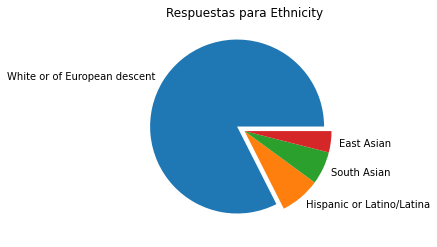




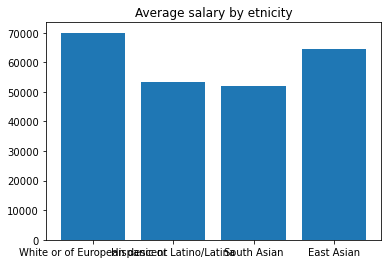
Datos originales Datos recortados 10%

Aunque existen valores extremos en casi todas las variables, solo en las etnias White or of European descent y Hispanic or Latino/Latina se refleja un recorte significativo de los mismos.

1. En la siguiente gráfica podemos observar que la etnia con la mayor cantidad de respuestas es ‘White or of European descent’ con 23,746.

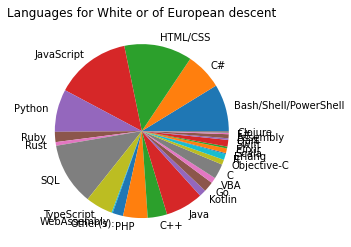


1. Haciendo un análisis de los salarios anuales promedio con respecto a las 4 etnias más comunes, se puede identificar que la etnia que tiende a tener los salarios más altos es ‘White or of European descent’ con un salario promedio de $69,989.75 y la etnia que tiende a tener los salaries mas bajos es ‘South Asian’ con un salario promedio de $51,966.29.



1. El análisis de los datos también nos muestra que los lenguajes de programación más populares y menos populares para cada una de las 4 etnias más comunes son los que se muestran a continuación:

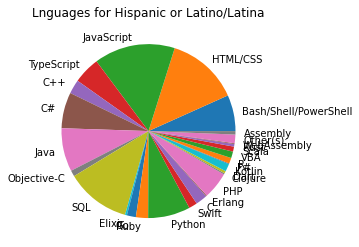
**Ethnicity = White or of European descent**



**El lenguaje más popular es: JavaScript.**

**El lenguaje menos popular es: Erlang**

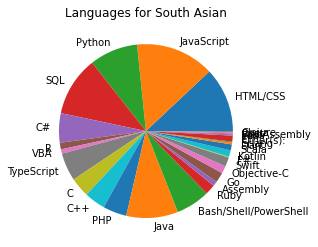
**Ethnicity = Hispanic or Latino/Latina**



**El lenguaje más popular es: JavaScript**

**El lenguaje menos popular es: F#**

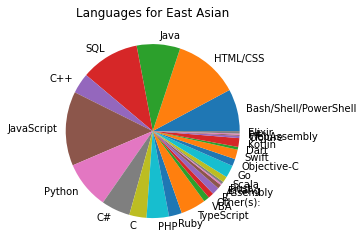
**Ethnicity = South Asian**



**El lenguaje más popular es: JavaScript**

**El lenguaje menos popular es: F#**

**Ethnicity = East Asian**

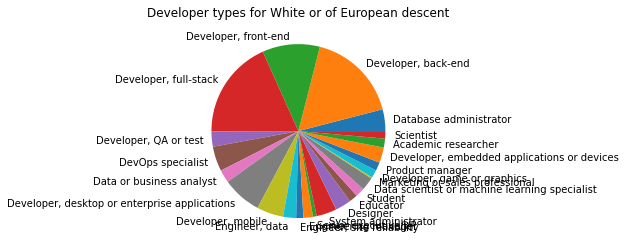


**El lenguaje más popular es: JavaScript**

**El lenguaje menos popular es: Web Assembly**

1. Al analizar los datos de la variable **DevType,** podemos encontrar cuales son los tipos de desarrollador más comunes para cada una de las 4 etnias más comunes:

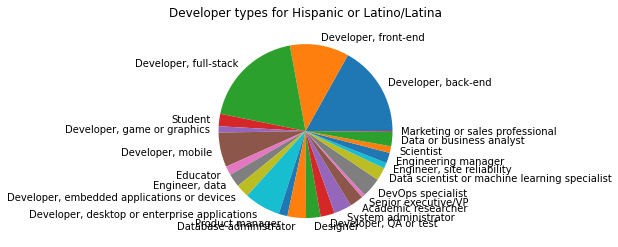
**Ethnicity = White or of European descent**



**El tipo de desarrollador más popular es: Developer, full-stack.**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

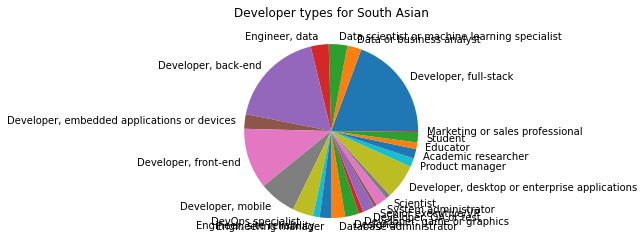
**Ethnicity = Hispanic or Latino/Latina**



**El tipo de desarrollador más popular es: Developer, full-stack**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

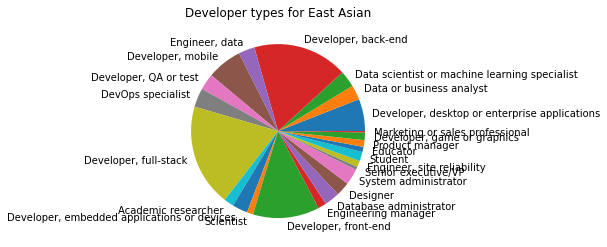
**Ethnicity = South Asian**



**El tipo de desarrollador más popular es: Developer, full-stack**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

**Ethnicity = East Asian**

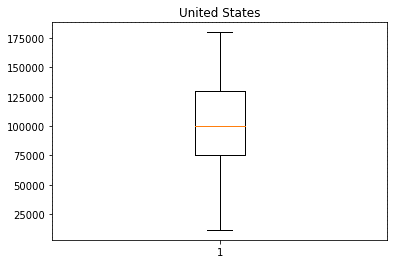


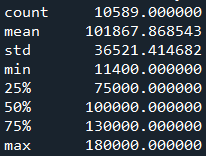
**El tipo de desarrollador más popular es: Developer, full-stack**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

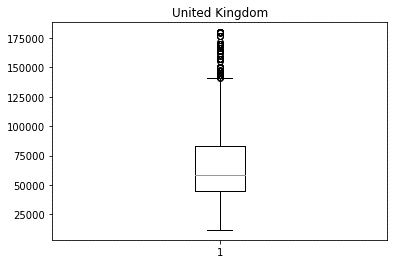
1. Al realizar un análisis preliminar (five-number summary, boxplot) de la variable para el salario anual (**ConvertedComp**) con respecto a los 4 países más populares podemos observar los siguientes resultados:

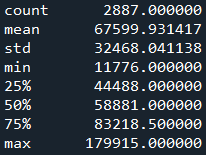
Para la variable **Country = United States**:



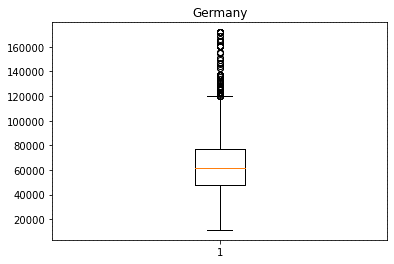


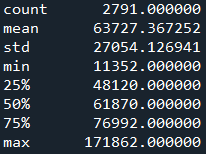
Para la variable **Country = United Kingdom**:



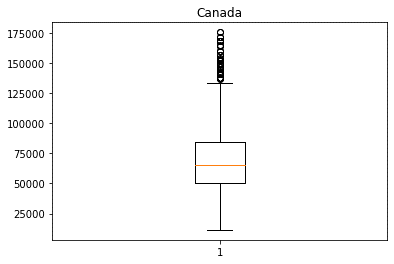


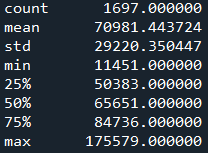
Para la variable **Country = Germany**:





Para la variable **Country = Canada**:

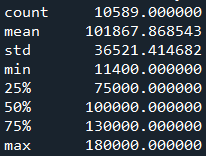
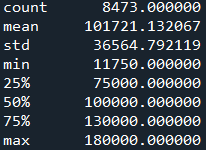




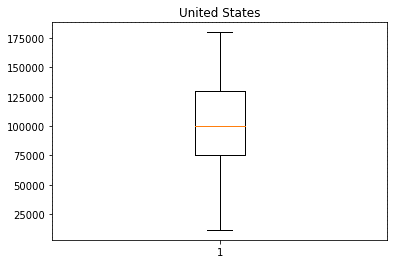
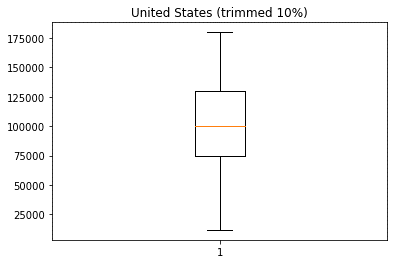
Haciendo un recorte de 10% a los datos, podemos observar cambios muy ligeros dentro del boxplot para todos los países excepto ‘United States’, esto representa la eliminación de valores extremos para la correcta interpretación de los datos.

A continuación, se muestran los datos originales y recortados al 10% para su comparación:

**Country = United States**

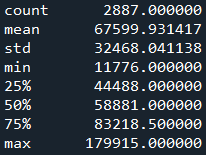
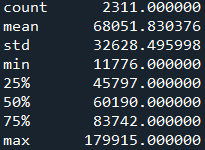


Datos originales Datos recortados 10%

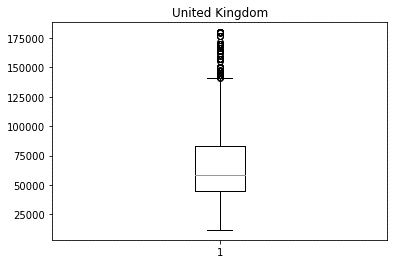
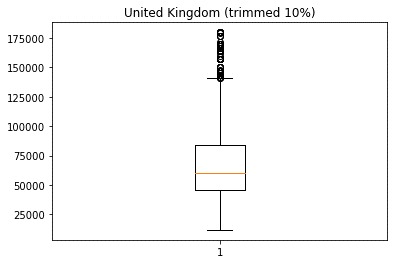


Datos originales Datos recortados 10%

**Country = United Kingdom**

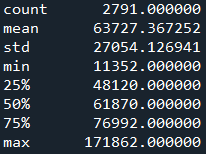
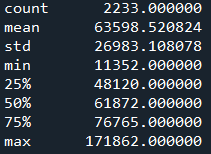


Datos originales Datos recortados 10%

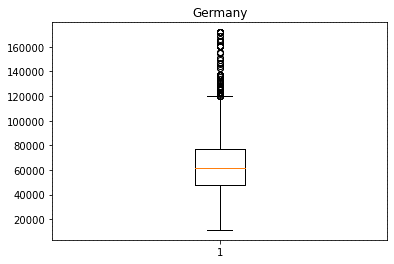


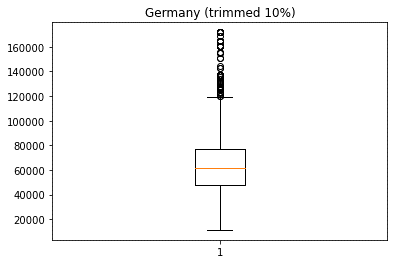
Datos originales Datos recortados 10%

**Country = Germany**



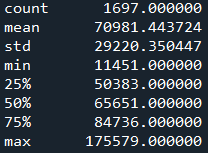
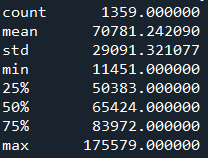
Datos originales Datos recortados 10%



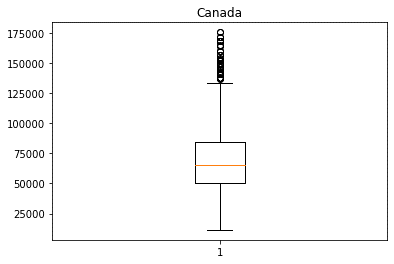


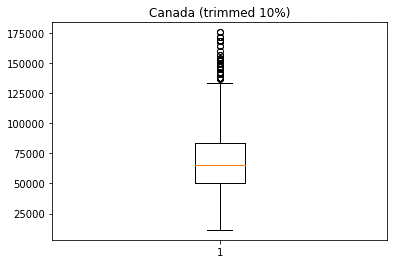
Datos originales Datos recortados 10%

**Country = Canada**



Datos originales Datos recortados 10%

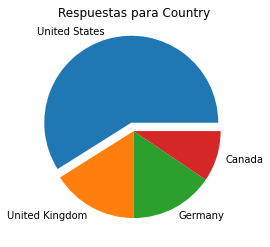




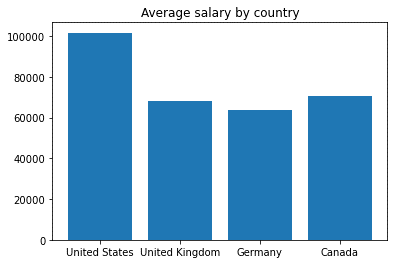
Datos originales Datos recortados 10%

Aunque se eliminaron algunos de los valores extremos dentro de las variables excepto ‘United States’, los valores extremos siguen representando un cierto desequilibrio dentro de las mismas. Es sobresaliente que la variable ‘United States’ no tiene valores extremos, lo que representa un equilibrio entre los datos recolectados sobre el salario anual para este país.

1. En la siguiente gráfica podemos observar que el país con la mayor cantidad de respuestas es ‘United States’ con 8,473.

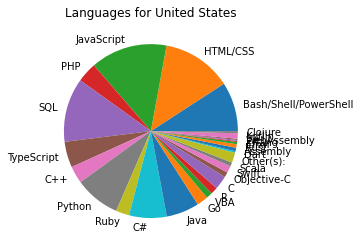


1. Haciendo un análisis de los salarios anuales promedio con respecto a los 4 países más comunes, se puede identificar que el país que tiende a tener los salarios más altos es ‘United States’ con un salario promedio de $101,721.13 y el país que tiende a tener los salarios más bajos es ‘Germany’ con un salario promedio de $63,598.52.



1. El análisis de los datos también nos muestra que los lenguajes de programación más populares y menos populares para cada uno de los 4 países más comunes son los que se muestran a continuación:

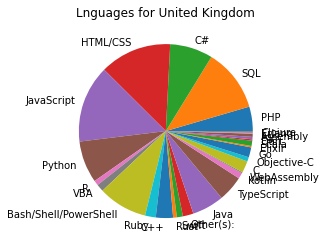
**Country = United States**



**El lenguaje más popular es: JavaScript.**

**El lenguaje menos popular es: Erlang**

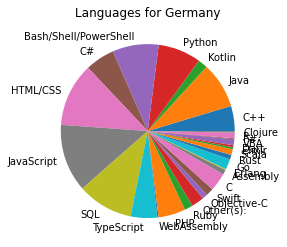
**Country = United Kingdom**



**El lenguaje más popular es: JavaScript**

**El lenguaje menos popular es: Erlang**

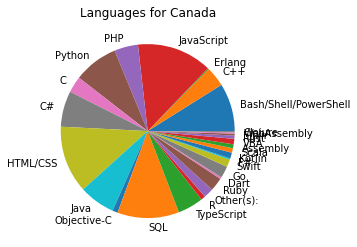
**Country = Germany**



**El lenguaje más popular es: JavaScript**

**El lenguaje menos popular es: F#**

**Country = Canada**

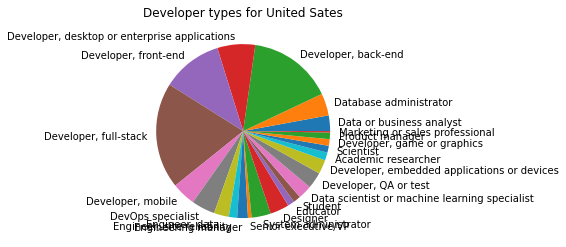


**El lenguaje más popular es: JavaScript**

**El lenguaje menos popular es: Erlang**

1. Al analizar los datos de la variable **DevType,** podemos encontrar cuales son los tipos de desarrollador más comunes para cada uno de los países más comunes:

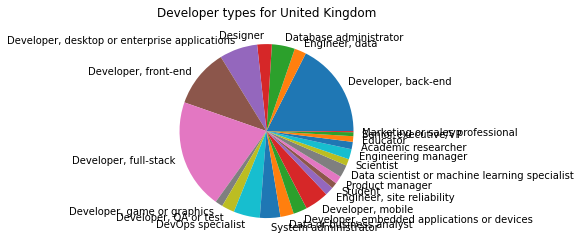
**Country = United States**



**El tipo de desarrollador más popular es: Developer, full-stack.**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

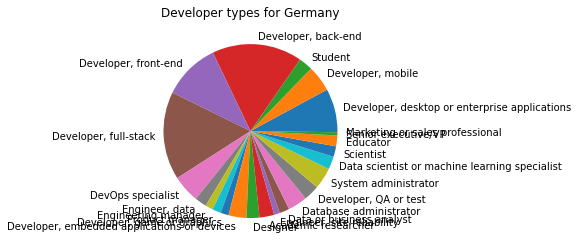
**Country = United Kingdom**



**El tipo de desarrollador más popular es: Developer, full-stack**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

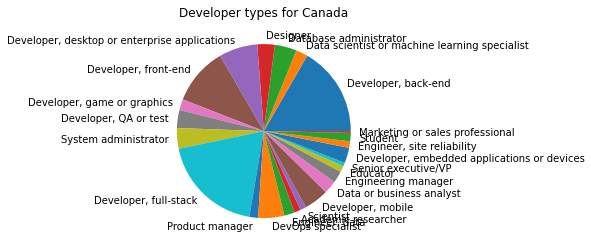
**Country = Germany**



**El tipo de desarrollador más popular es: Developer, back-end**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

**Country = Canada**

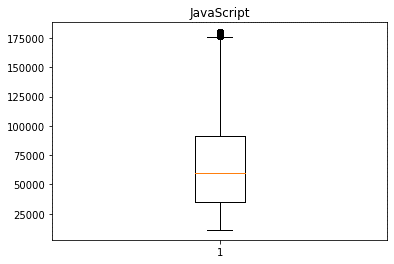


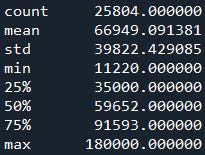
**El tipo de desarrollador más popular es: Developer, full-stack**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

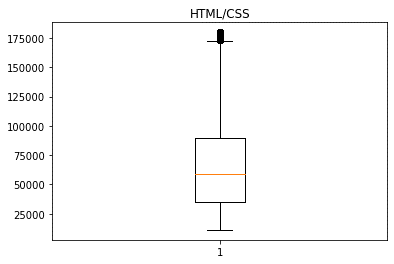
1. Al realizar un análisis preliminar (five-number summary, boxplot) de la variable para el salario anual (**ConvertedComp**) con respecto a los 4 lenguajes de programación más populares podemos observar los siguientes resultados:

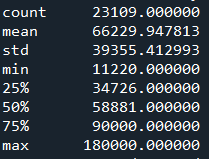
Para la variable **LanguageWorkedWith = JavaScript:**



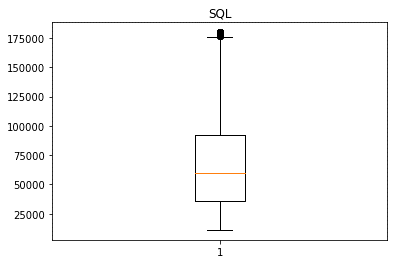


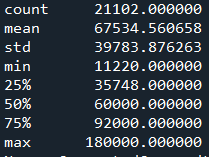
Para la variable **LanguageWorkedWith = HTML/CSS**:



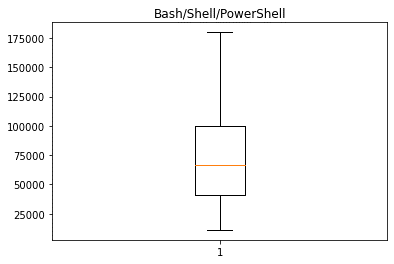


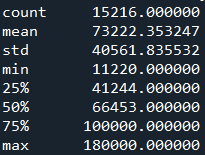
Para la variable **LanguageWorkedWith = SQL**:





Para la variable **LanguageWorkedWith = Bash/Shell/PowerShell**:

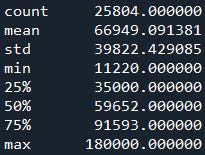
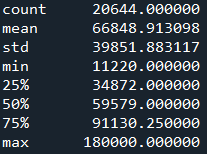




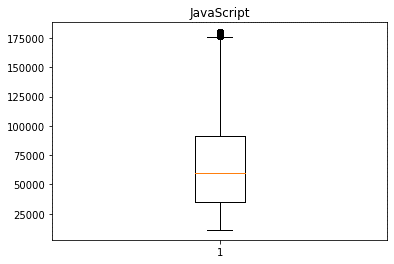
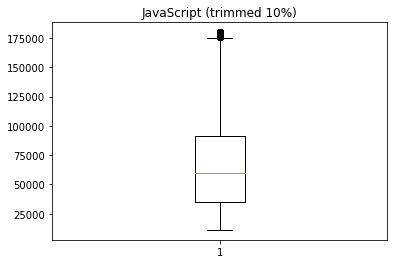
Haciendo un recorte de 10% a los datos, podemos observar cambios muy ligeros dentro del boxplot, lo cual significa que aunque existan valores extremos, no representan una inconveniencia para la correcta interpretación de los datos.

A continuación, se muestran los datos originales y recortados al 10% para su comparación:

**LanguageWorkedWith = JavaScript**

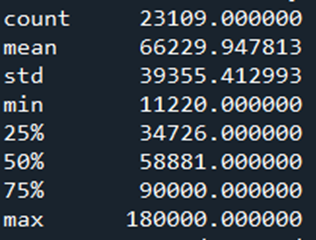
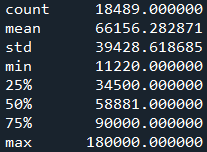


Datos originales Datos recortados 10%

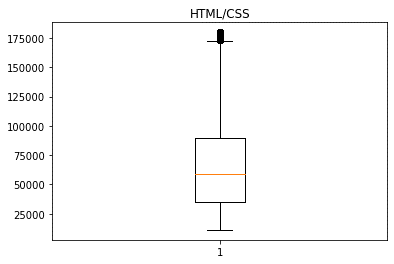
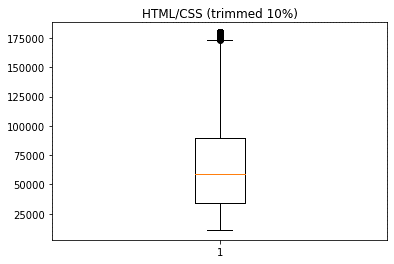


Datos originales Datos recortados 10%

**LanguageWorkedWith = HTML/CSS**

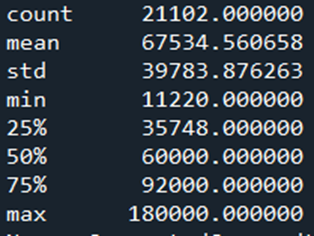
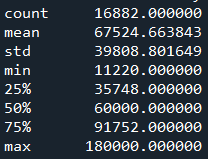


Datos originales Datos recortados 10%

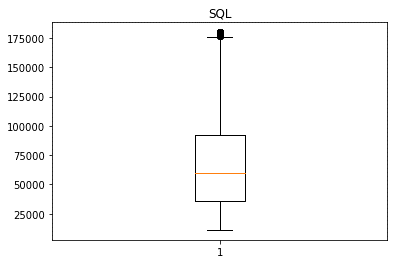


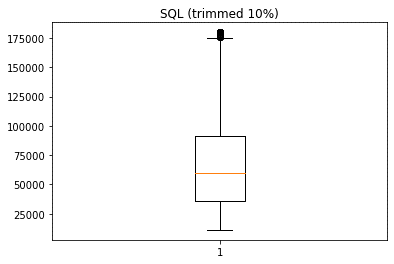
Datos originales Datos recortados 10%

**LanguageWorkedWith = SQL**



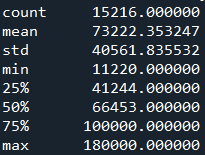
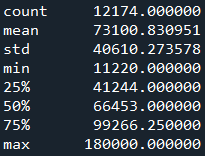
Datos originales Datos recortados 10%



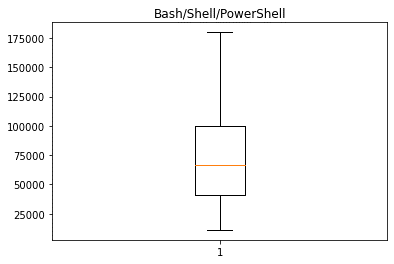


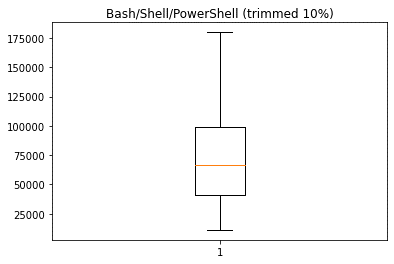
Datos originales Datos recortados 10%

**LanguageWorkedWith = Bash/Shell/PowerShell**



Datos originales Datos recortados 10%

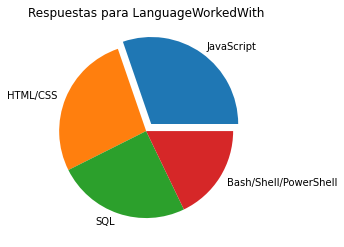




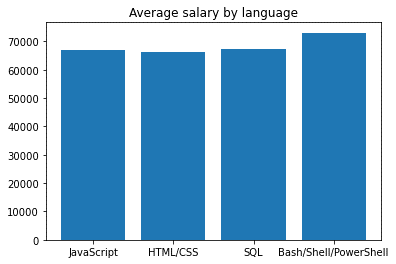
Datos originales Datos recortados 10%

Después de hacer el recorte del 10% de los datos, podemos observar que la diferencia entre las estadísticas es muy pequeña. Aunque existen valores extremos en la mayoría de las variables, la dispersión de los datos se mantiene entre los datos originales y los recortados.

1. En la siguiente gráfica podemos observar que el lenguaje de programación con la mayor cantidad de respuestas es ‘JavaScript’ con 20,644.

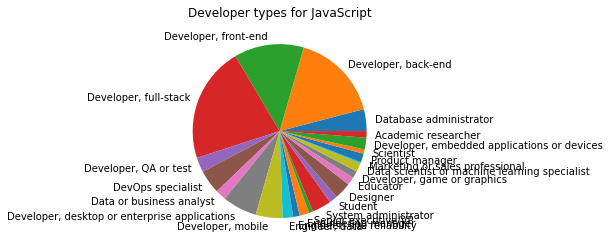


1. Haciendo un análisis de los salarios anuales promedio con respecto a los 4 lenguajes de programación más comunes, se puede identificar que el lenguaje que tiende a tener los salarios más altos es ‘Bash/Shell/PowerShell’ con un salario promedio de $73,100.83 y el lenguaje que tiende a tener los salarios más bajos es ‘HTML/CSS’ con un salario promedio de $66,156.28.



1. Al analizar los datos de la variable **DevType,** podemos encontrar cuales son los tipos de desarrollador más comunes para cada uno de los lenguajes de programación más comunes:

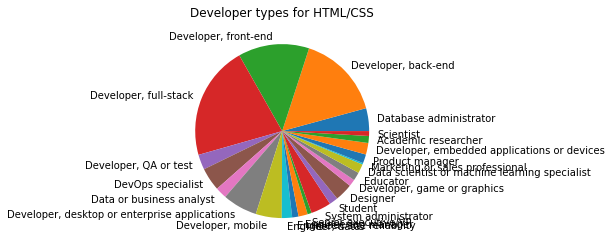
**LanguageWorkedWith = JavaScript**



**El tipo de desarrollador más popular es: Developer, full-stack.**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

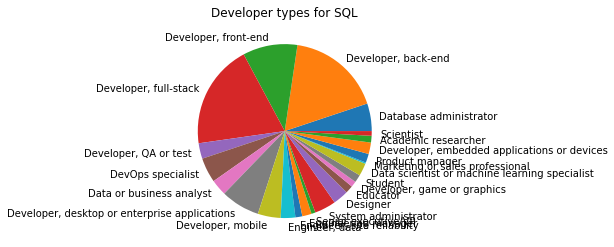
**LanguageWorkedWith = HTML/CSS**



**El tipo de desarrollador más popular es: Developer, full-stack**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

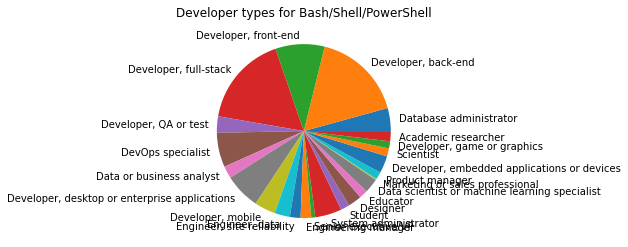
**LanguageWorkedWith = SQL**



**El tipo de desarrollador más popular es: Developer, full-stack**

**El tipo de desarrollador menos popular es: Marketing or sales professional**

**LanguageWorkedWith = Bash/Shell/PowerShell**



**El tipo de desarrollador más popular es: Developer, full-stack**

**El tipo de desarrollador menos popular es: Marketing or sales professional**