MINERÍA DE DATOS

Práctica

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.

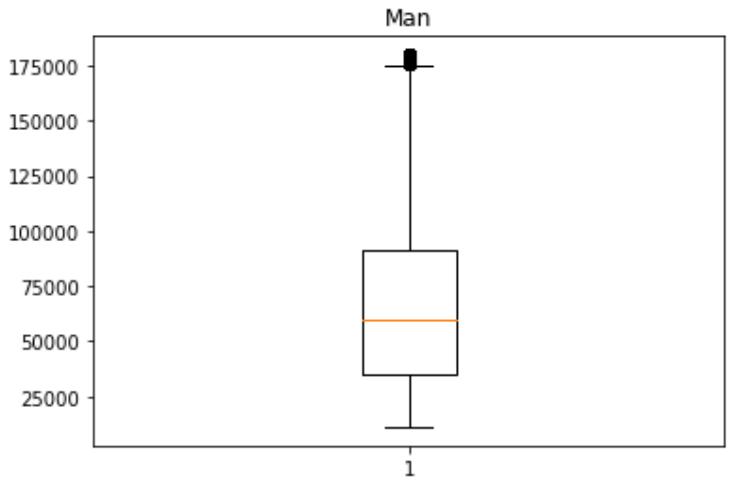
Iván Daniel Hernández Rocha

id.hernandezrocha@ugto.mx

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:

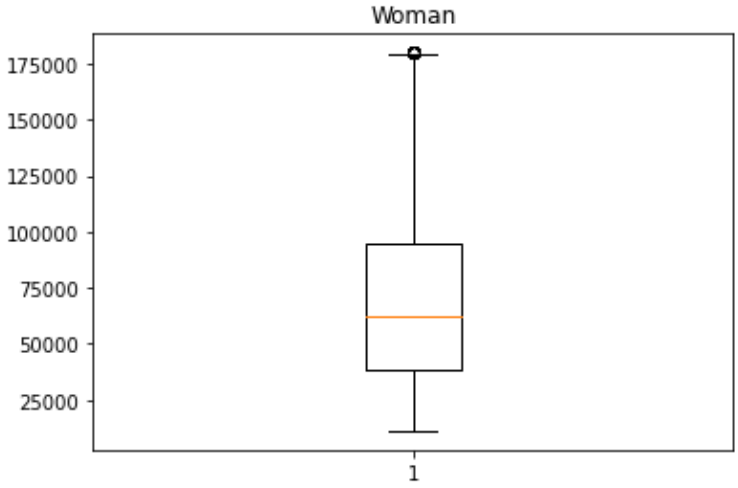
   Para la variable **Gender = Man**:

count	33826.000000
mean	66806.032253
std	39916.808802
min	11220.000000
25%	34791.000000
50%	59579.000000
75%	91000.000000
max	180000.000000



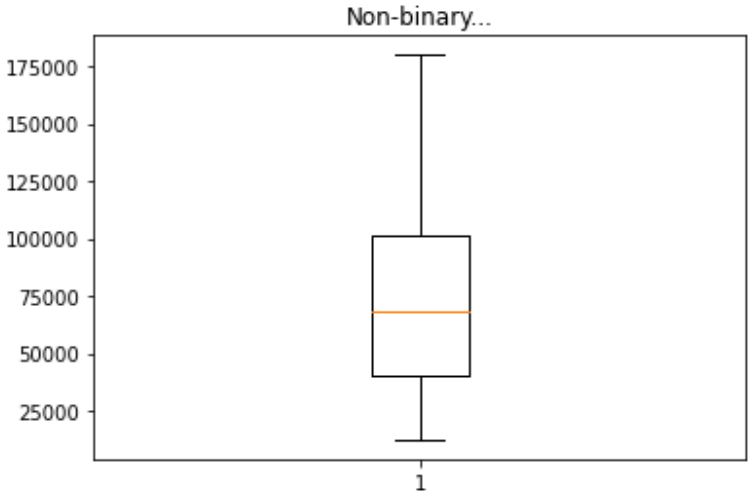
Para la variable **Gender = Woman**:

count	2714.000000
mean	69191.874724
std	39344.494582
min	11220.000000
25%	38496.000000
50%	62468.000000
75%	94915.000000
max	180000.000000



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

count	390.000000
mean	74620.466667
std	41465.024132
min	12372.000000
25%	40146.750000
50%	68232.500000
75%	101406.500000
max	180000.000000



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:

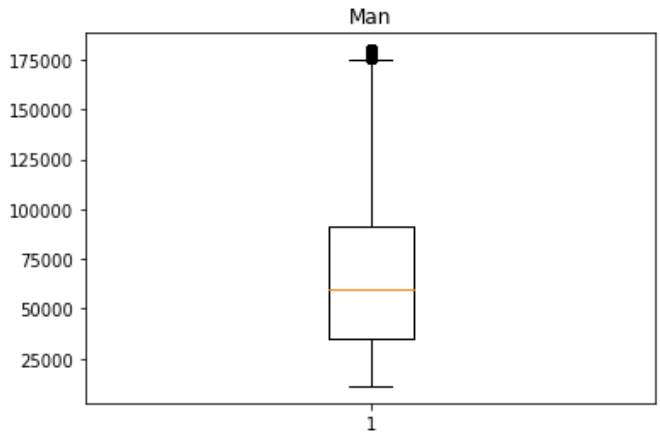
Gender = Man

count	33826.000000
mean	66806.032253
std	39916.808802
min	11220.000000
25%	34791.000000
50%	59579.000000
75%	91000.000000
max	180000.000000

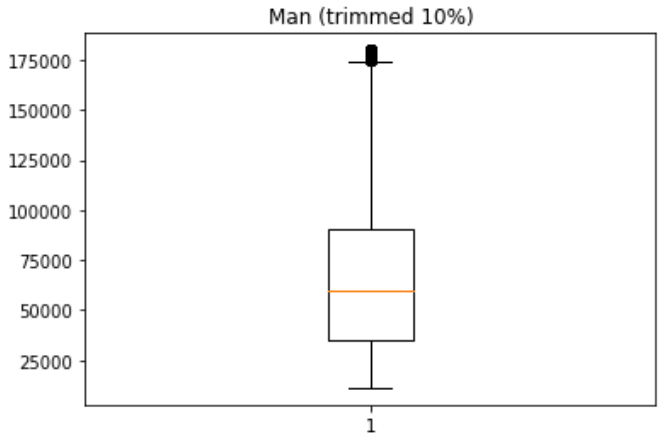
Datos originales

count	27062.000000
mean	66791.977090
std	39934.839005
min	11220.000000
25%	34791.000000
50%	59579.000000
75%	90662.000000
max	180000.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

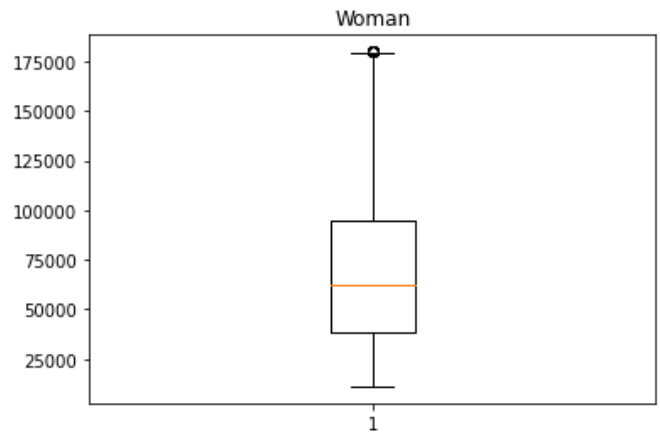
Gender = Woman

count	2714.000000
mean	69191.874724
std	39344.494582
min	11220.000000
25%	38496.000000
50%	62468.000000
75%	94915.000000
max	180000.000000

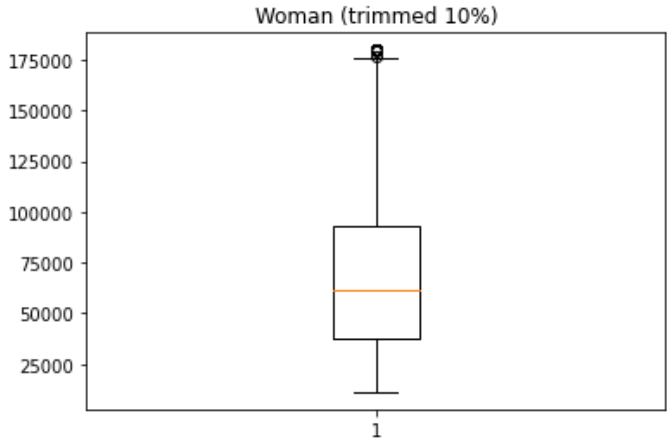
Datos originales

count	2172.000000
mean	68755.049263
std	39275.638967
min	11220.000000
25%	37986.500000
50%	61872.000000
75%	93250.000000
max	180000.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

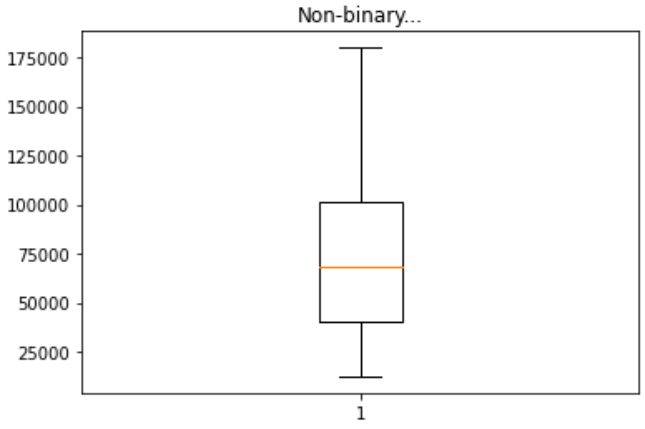
Gender = Non-binary, genderqueer, or gender non-conforming

count	390.000000
mean	74620.466667
std	41465.024132
min	12372.000000
25%	40146.750000
50%	68232.500000
75%	101406.500000
max	180000.000000

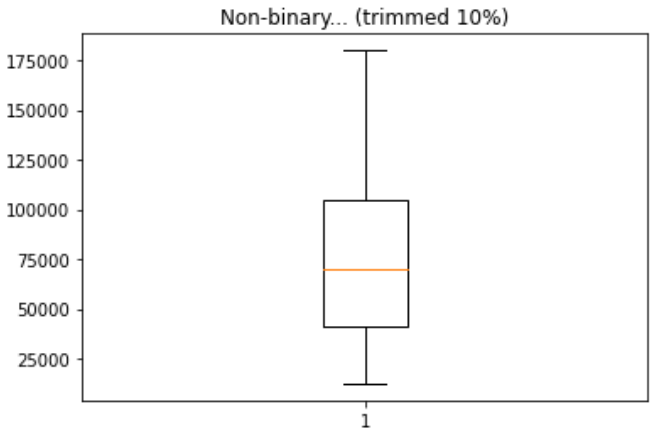
Datos originales

count	312.000000
mean	76178.586538
std	41601.170421
min	12372.000000
25%	41244.000000
50%	70000.000000
75%	105000.000000
max	180000.000000

Datos recortados 10%



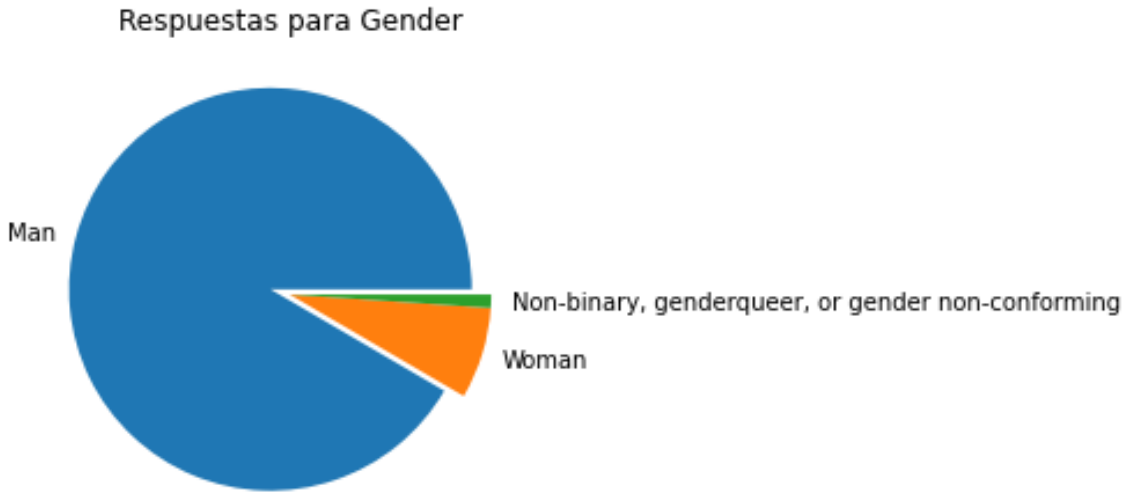
Datos originales



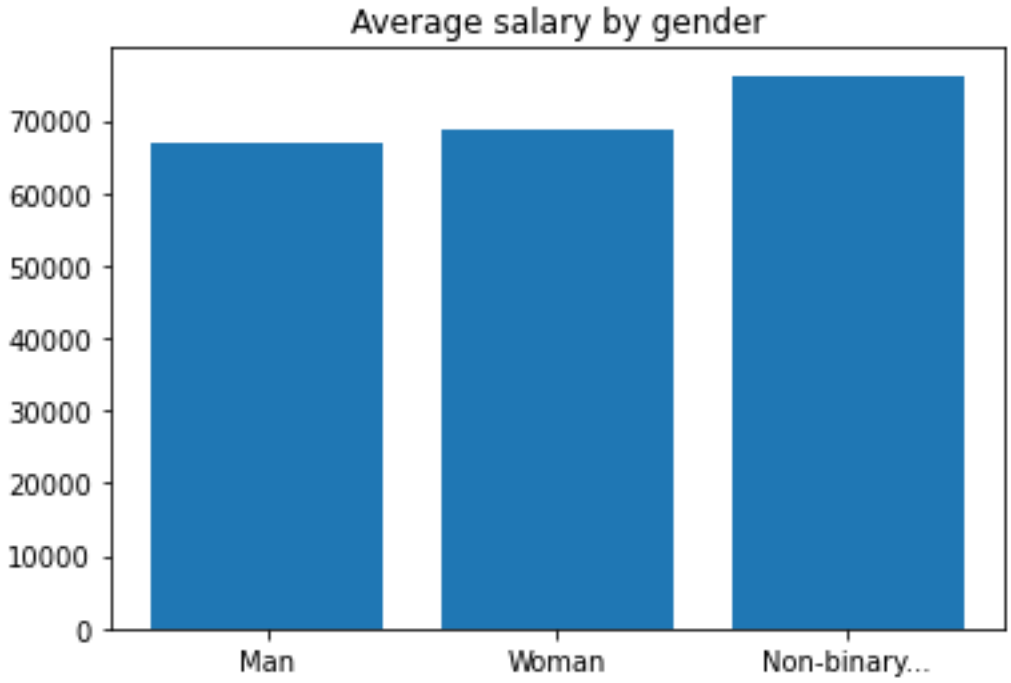
Datos recortados 10%

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.

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

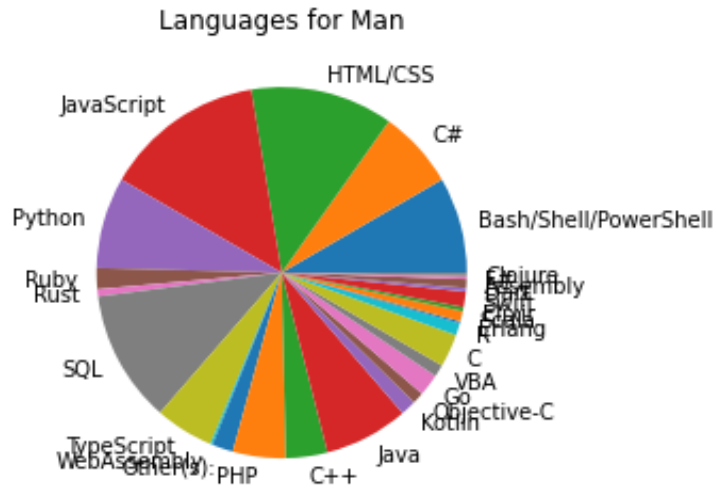


b. Haciendo un análisis de los salarios anuales promedio, podemos observar que el género que tiende a tener los salarios más altos es ‘Non-binary, genderqueer, or gender non-conforming’, con un salario promedio de $76,178.59 y a su vez, el género que tiende a tener los salarios más bajos es ‘Man’ con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



c. 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:

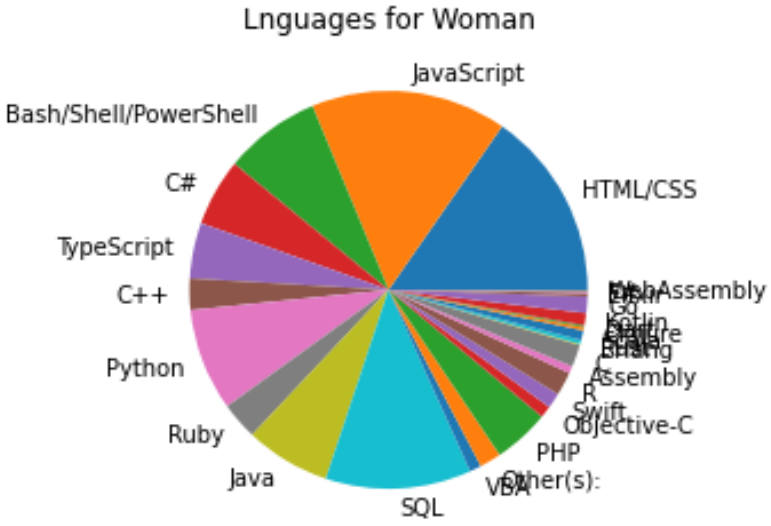
Gender = Man



El lenguaje más popular es: JavaScript.

El lenguaje menos popular es Web Assembly

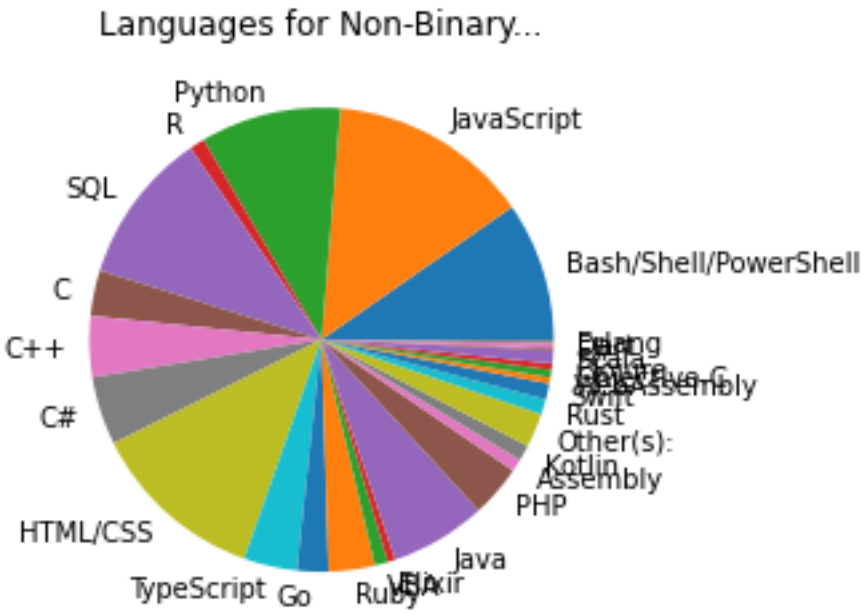
Gender = Woman



El lenguaje más popular es: JavaScript

El lenguaje menos popular es: Erlang

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

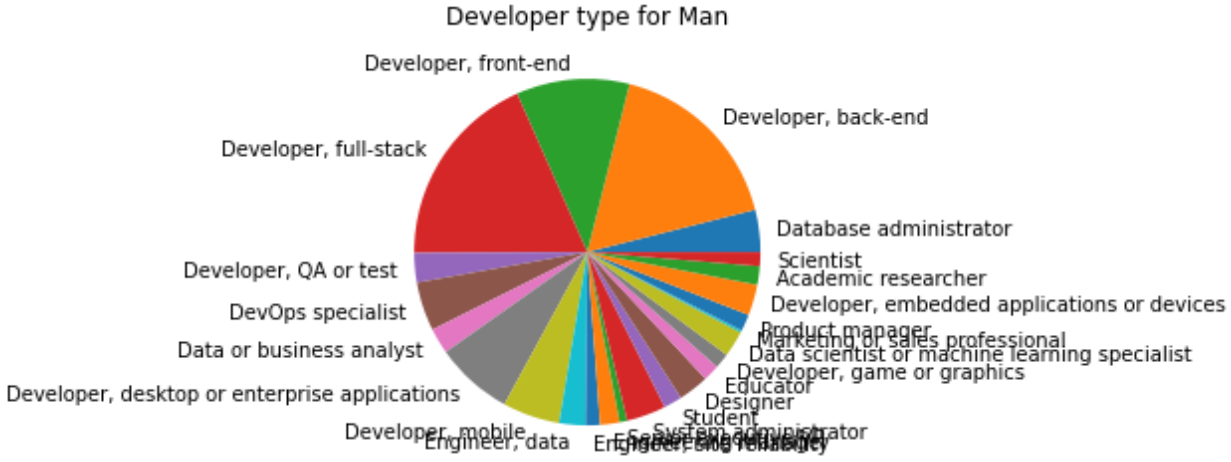


El lenguaje más popular es: JavaScript

El lenguaje menos popular es: F#

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

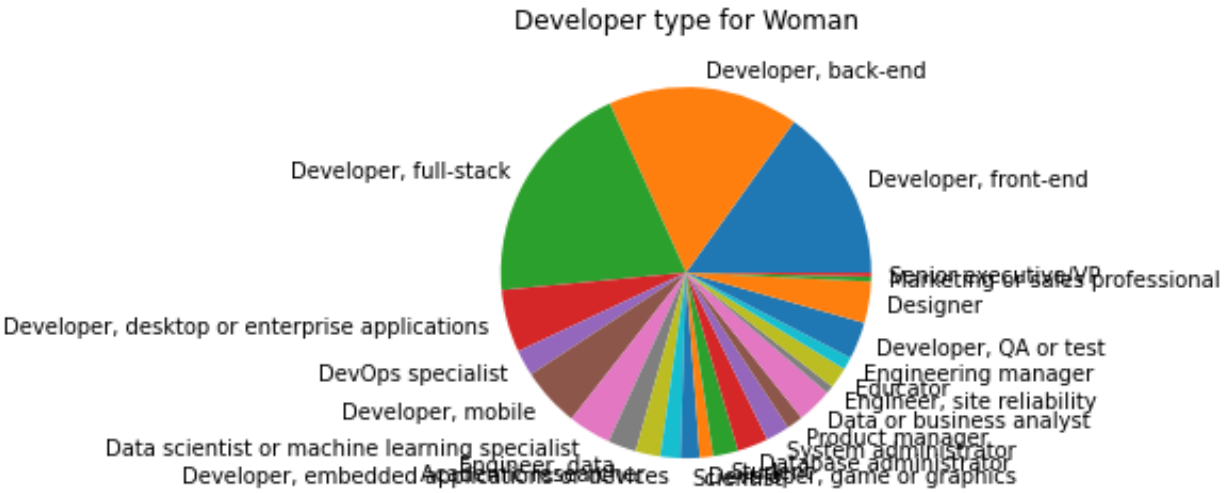
Gender = Man



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

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

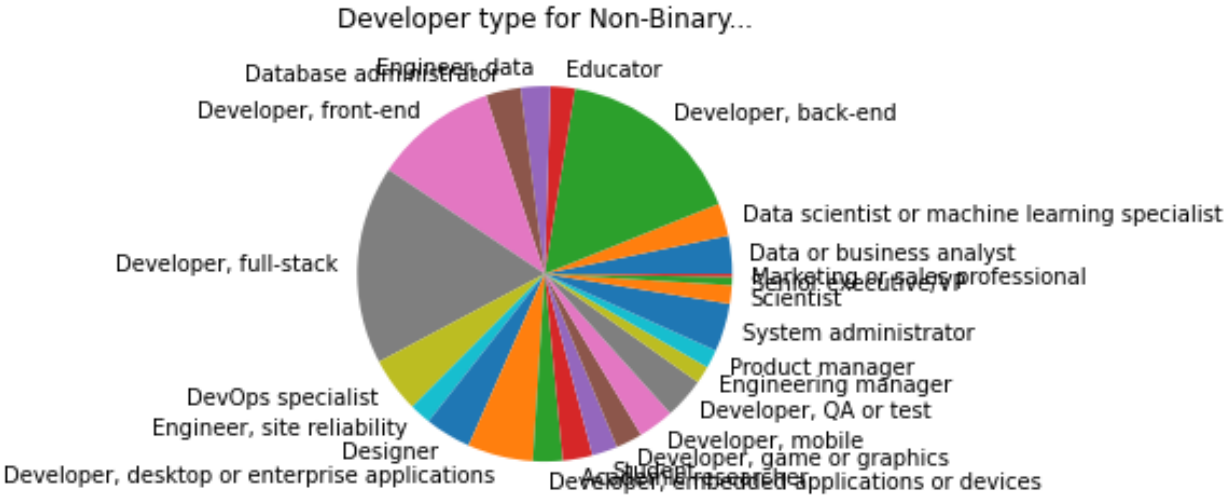
Gender = Woman



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



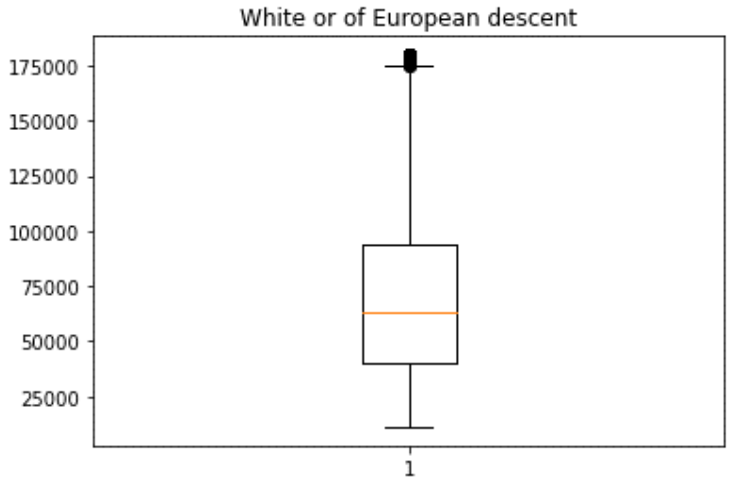
El tipo de desarrollador más popular es: Developer, full-stack

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

2. 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:

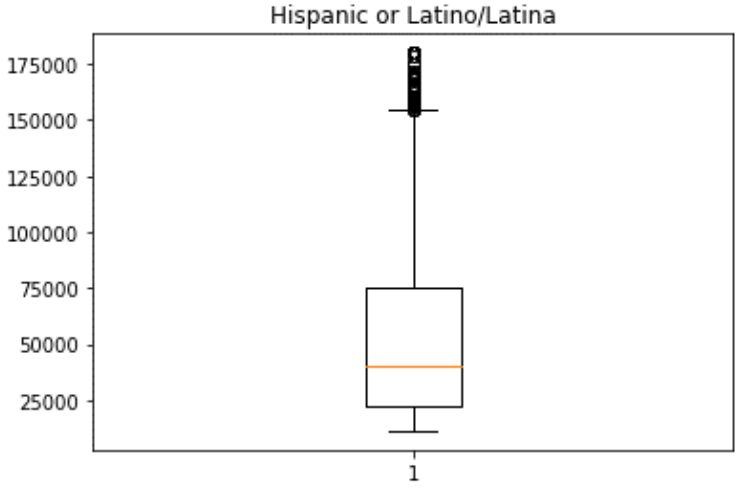
   Para la variable **Ethnicity = White or of European descent**:

count	29682.000000
mean	70032.857995
std	39188.106621
min	11220.000000
25%	39879.750000
50%	63016.000000
75%	94000.000000
max	180000.000000



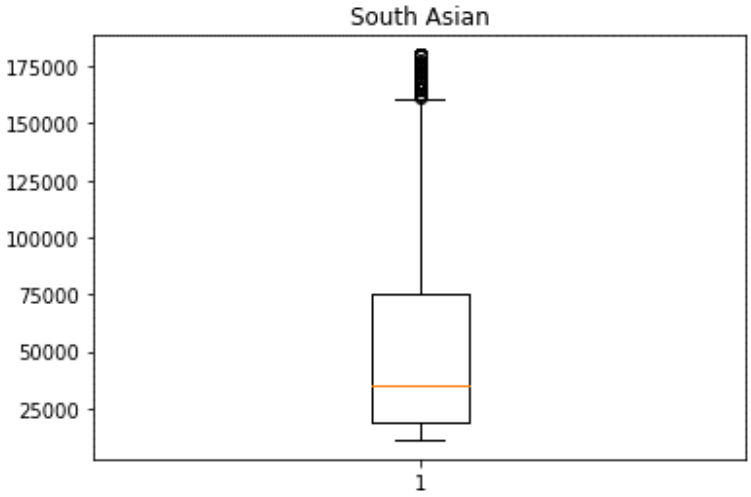
Para la variable **Ethnicity = Hispanic or Latino/Latina**:

count	2688.000000
mean	53398.584449
std	39560.737279
min	11268.000000
25%	21999.000000
50%	40050.000000
75%	75000.000000
max	180000.000000



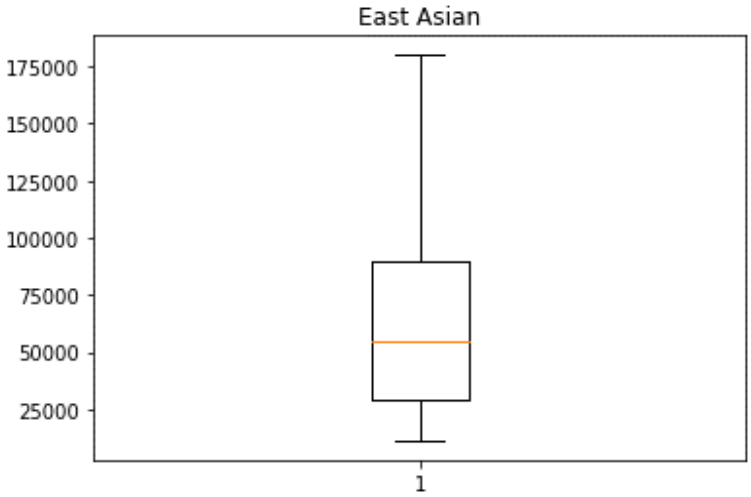
Para la variable **Ethnicity = South Asian**:

count	2201.000000
mean	52071.388005
std	41305.174184
min	11256.000000
25%	18576.000000
50%	34982.000000
75%	75564.000000
max	180000.000000



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

count	1419.000000
mean	63973.908386
std	41381.298305
min	11268.000000
25%	29436.500000
50%	54792.000000
75%	90000.000000
max	180000.000000



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:

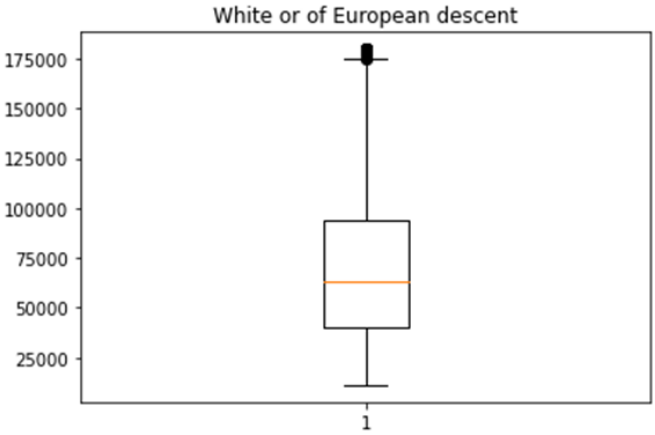
Ethnicity = White or of European descent

count	29682.000000
mean	70032.857995
std	39188.106621
min	11220.000000
25%	39879.750000
50%	63016.000000
75%	94000.000000
max	180000.000000

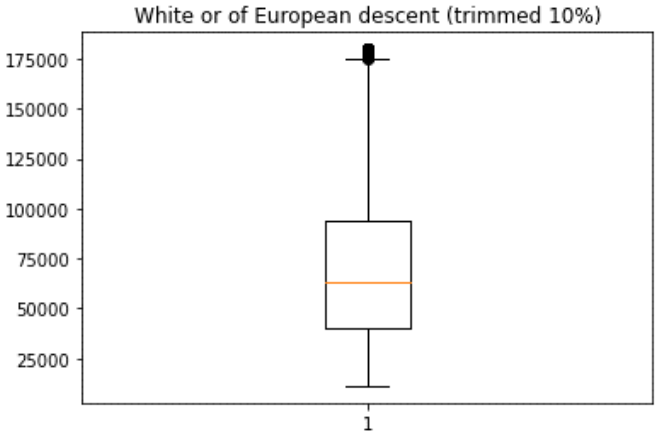
Datos originales

count	23746.000000
mean	69989.751706
std	39211.858750
min	11220.000000
25%	39908.000000
50%	63016.000000
75%	93951.000000
max	180000.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

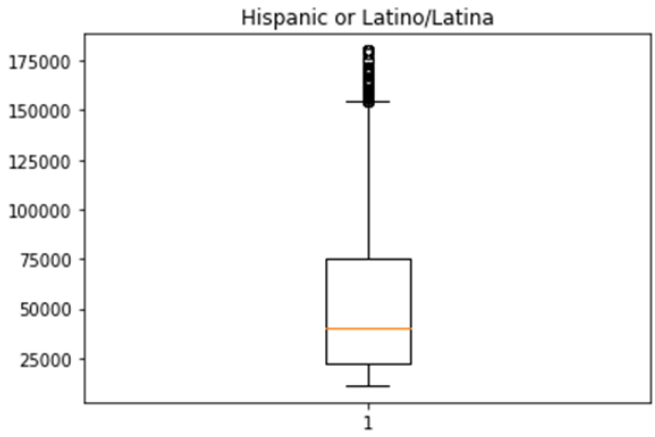
Ethnicity = Hispanic or Latino/Latina

count	2688.000000
mean	53398.584449
std	39560.737279
min	11268.000000
25%	21999.000000
50%	40050.000000
75%	75000.000000
max	180000.000000

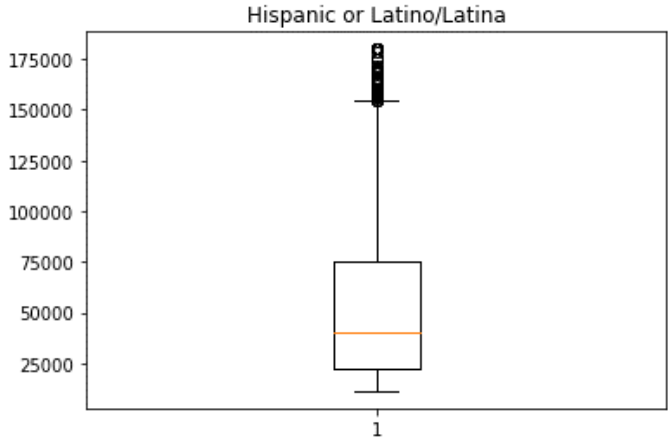
Datos originales

count	2152.000000
mean	53489.241636
std	39782.840893
min	11268.000000
25%	22020.000000
50%	40000.000000
75%	74474.000000
max	180000.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

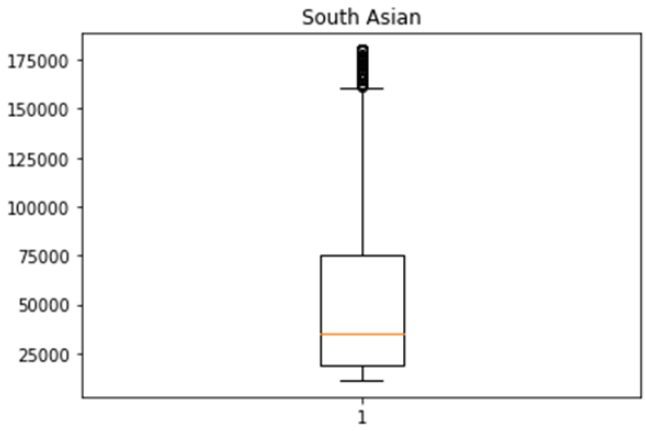
Ethnicity = South Asian

count	2201.000000
mean	52071.388005
std	41305.174184
min	11256.000000
25%	18576.000000
50%	34982.000000
75%	75564.000000
max	180000.000000

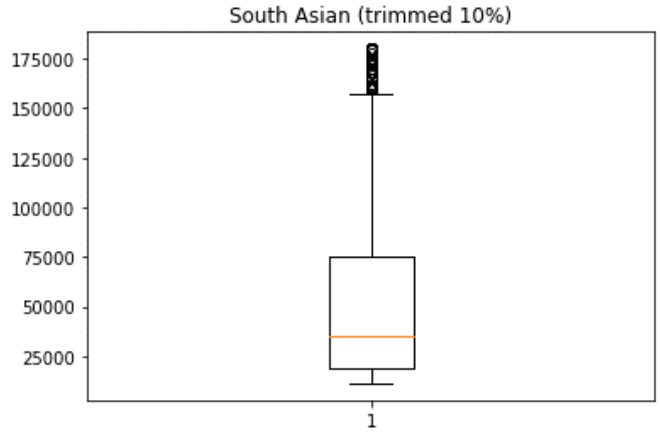
Datos originales

count	1761.000000
mean	51966.294719
std	41367.627394
min	11256.000000
25%	18576.000000
50%	34982.000000
75%	75000.000000
max	180000.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

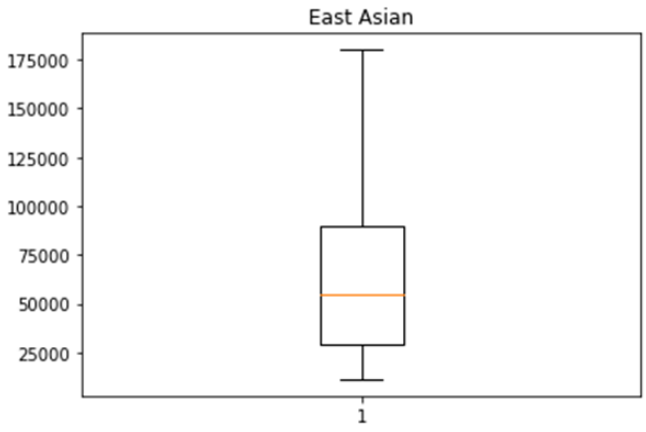
Ethnicity = East Asian

count	1419.000000
mean	63973.908386
std	41381.298305
min	11268.000000
25%	29436.500000
50%	54792.000000
75%	90000.000000
max	180000.000000

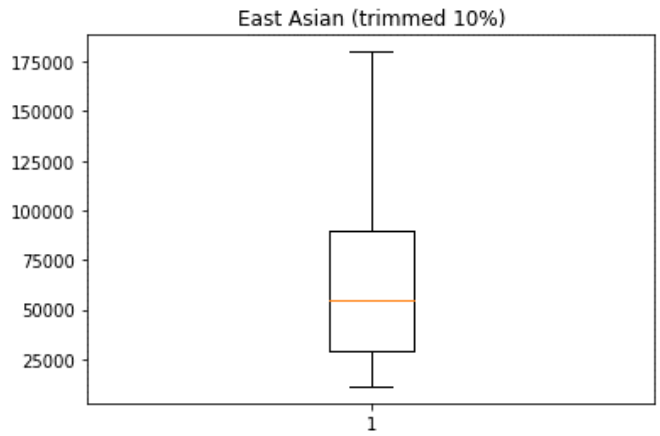
Datos originales

count	1137.000000
mean	64516.813544
std	41709.300699
min	11268.000000
25%	29328.000000
50%	54797.000000
75%	90000.000000
max	180000.000000

Datos recortados 10%



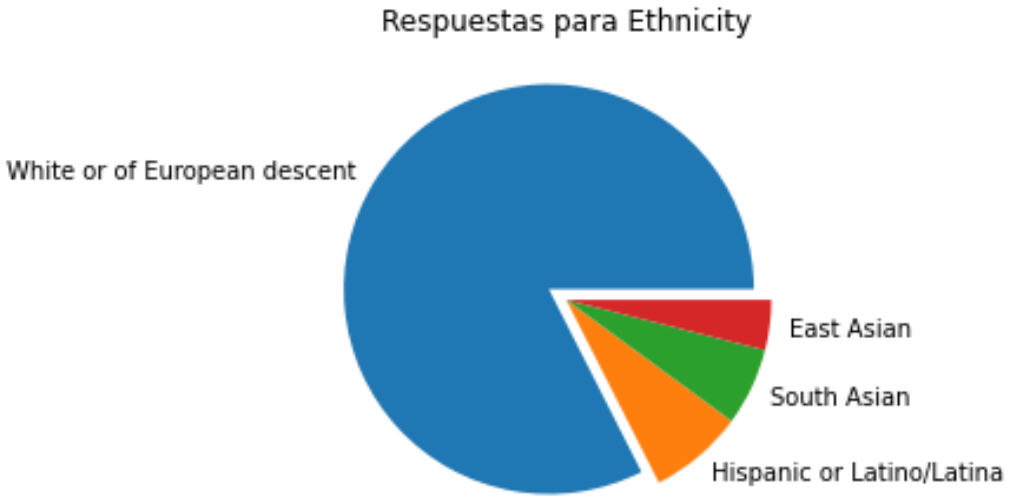
Datos originales



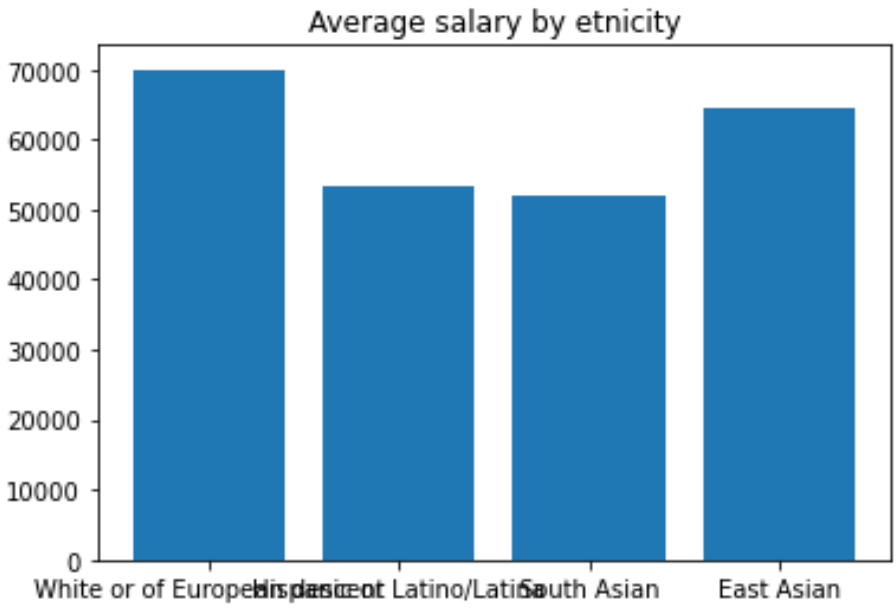
Datos recortados 10%

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.

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

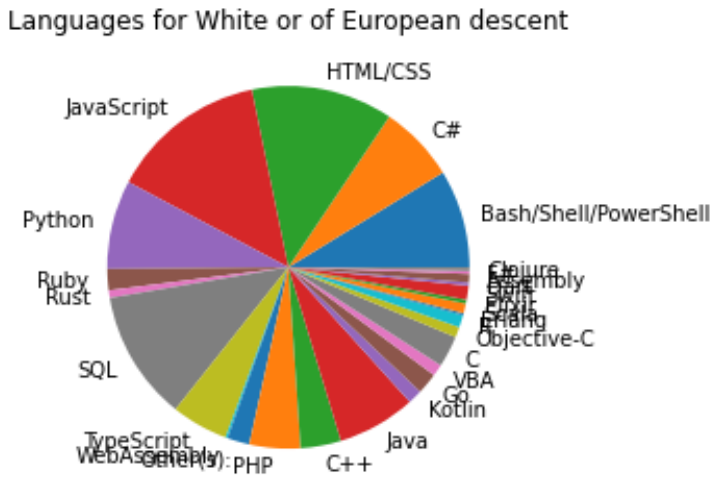


b. Haciendo un análisis de los salarios anuales promedio con respecto a las 4 etnias más comunes, se puede identificar que la etnia que tiende a tener los salarios más altos es ‘White or of European descent’ con un salario promedio de $69,989.75 y la etnia que tiende a tener los salaries mas bajos es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



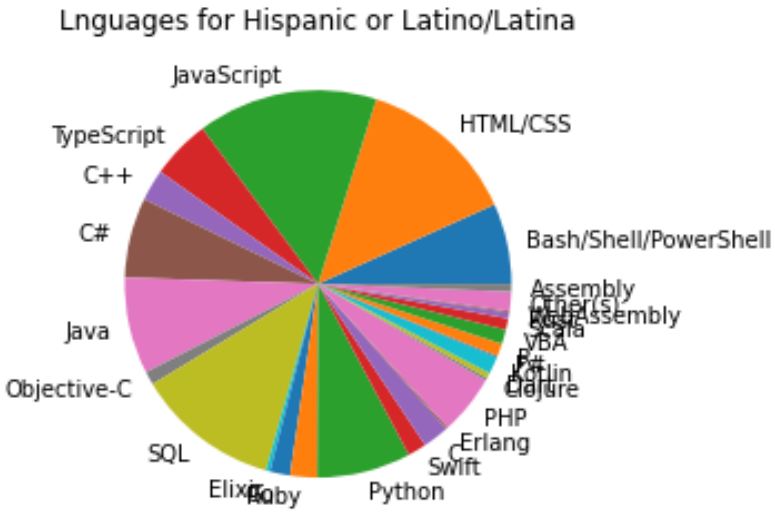
c. El análisis de los datos también nos muestra que los lenguajes de programación más populares y menos populares para cada una de las 4 etnias más comunes son los que se muestran a continuación:

Ethnicity = White or of European descent



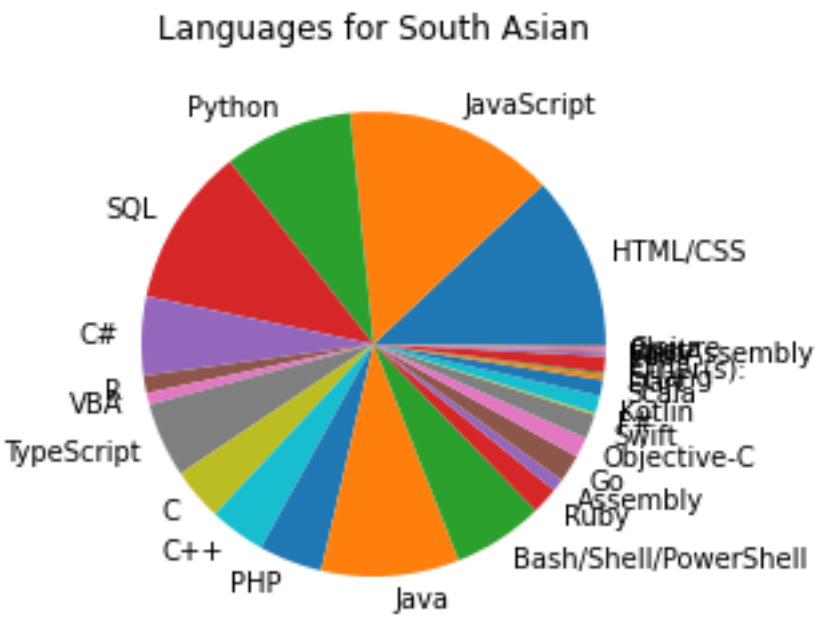
El lenguaje más popular es: JavaScript.
El lenguaje menos popular es: Erlang

Ethnicity = Hispanic or Latino/Latina



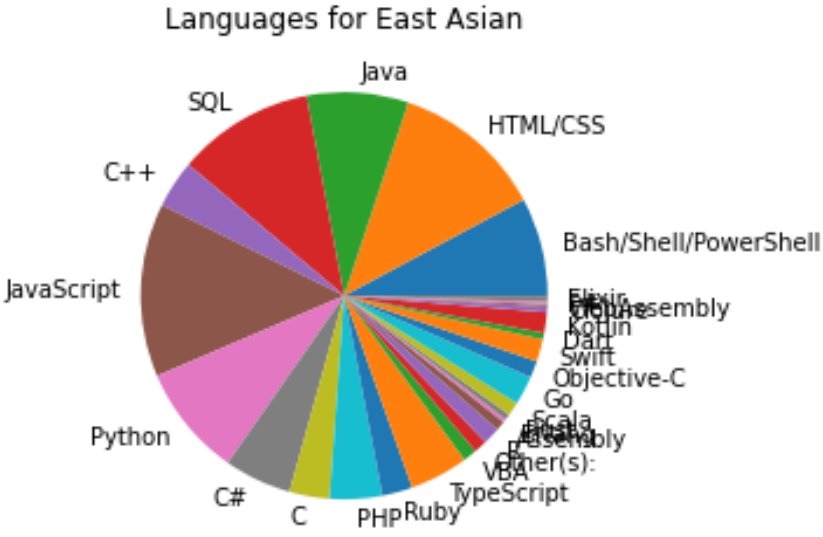
El lenguaje más popular es: JavaScript
El lenguaje menos popular es: F#

Ethnicity = South Asian



El lenguaje más popular es: JavaScript
El lenguaje menos popular es: F#

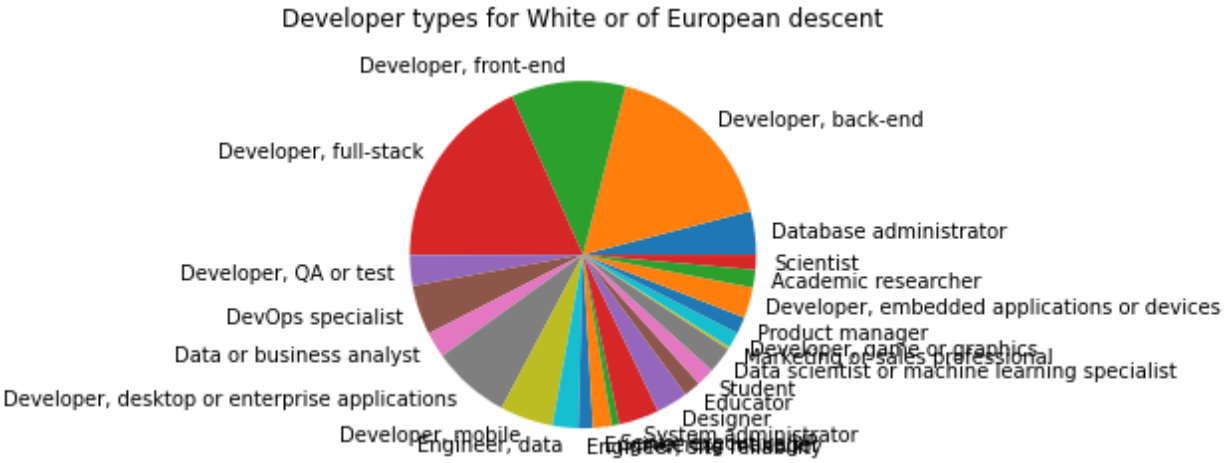
Ethnicity = East Asian



El lenguaje más popular es: JavaScript
El lenguaje menos popular es: Web Assembly

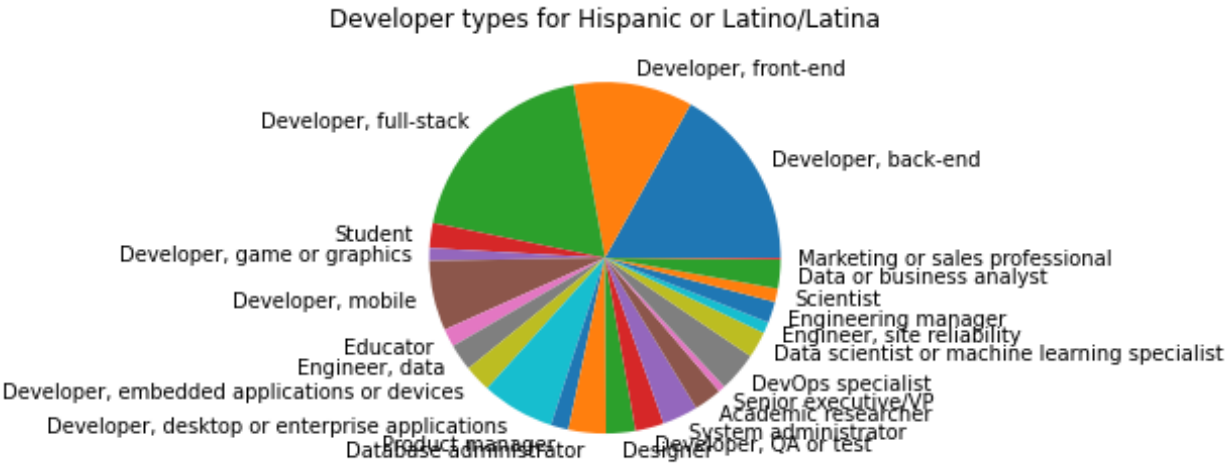
d. Al analizar los datos de la variable **DevType,** podemos encontrar cuales son los tipos de desarrollador más comunes para cada una de las 4 etnias más comunes:

Ethnicity = White or of European descent



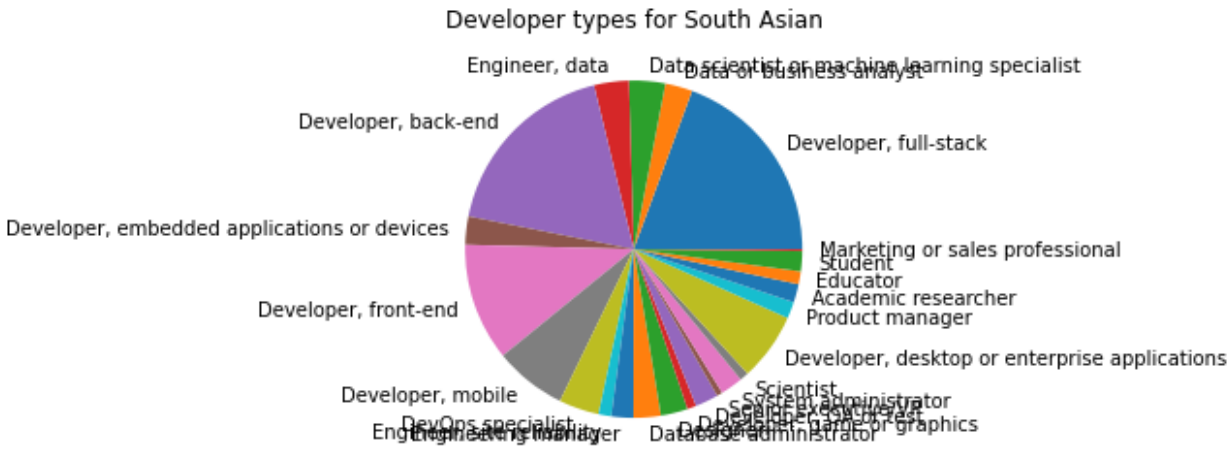
El tipo de desarrollador más popular es: Developer, full-stack.
El tipo de desarrollador menos popular es: Marketing or sales professional

Ethnicity = Hispanic or Latino/Latina



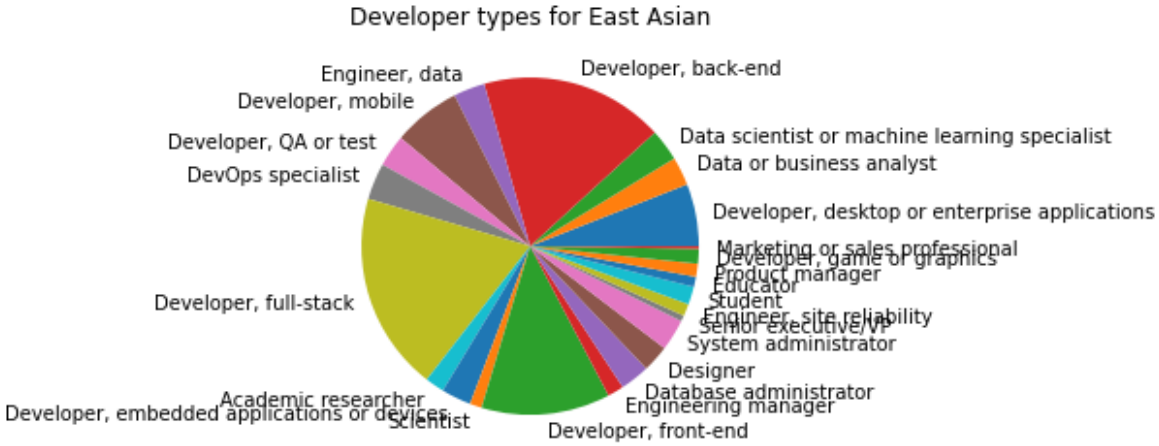
El tipo de desarrollador más popular es: Developer, full-stack
El tipo de desarrollador menos popular es: Marketing or sales professional

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

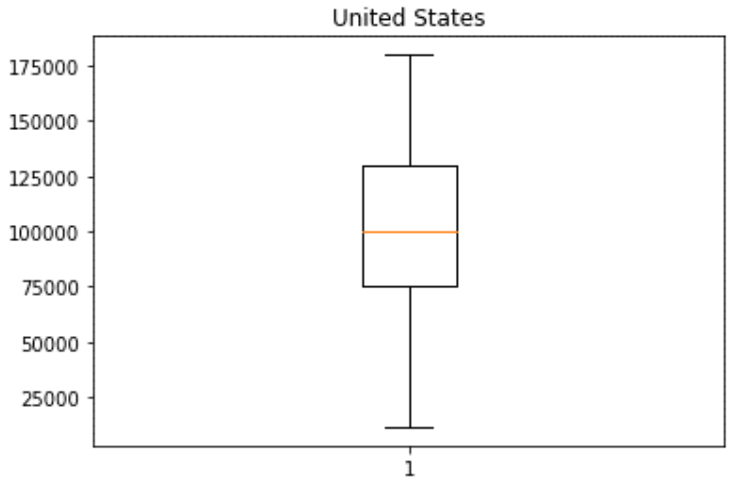


El tipo de desarrollador más popular es: Developer, full-stack
El tipo de desarrollador menos popular es: Marketing or sales professional

3. 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:

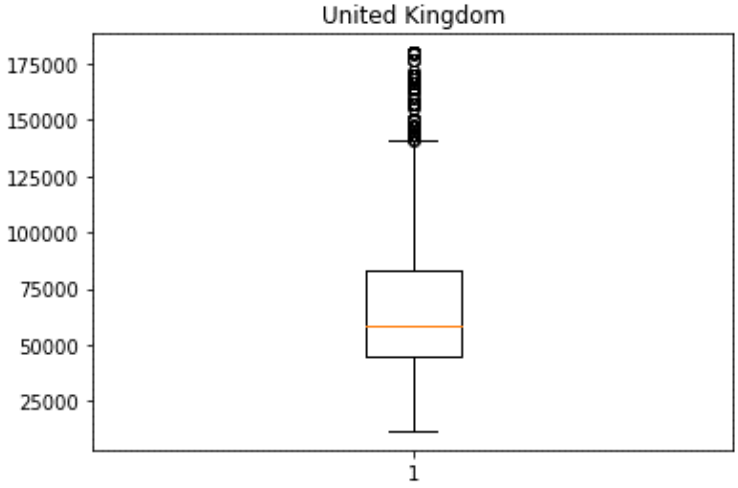
   Para la variable **Country = United States**:

count	10589.000000
mean	101867.868543
std	36521.414682
min	11400.000000
25%	75000.000000
50%	100000.000000
75%	130000.000000
max	180000.000000



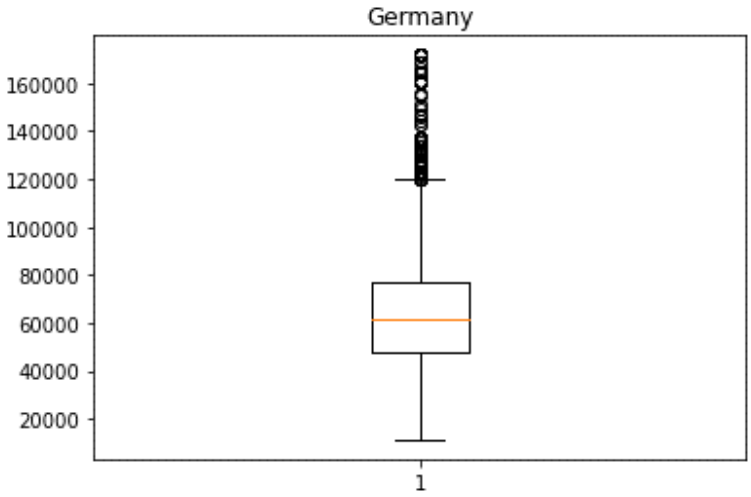
Para la variable **Country = United Kingdom**:

count	2887.000000
mean	67599.931417
std	32468.041138
min	11776.000000
25%	44488.000000
50%	58881.000000
75%	83218.500000
max	179915.000000



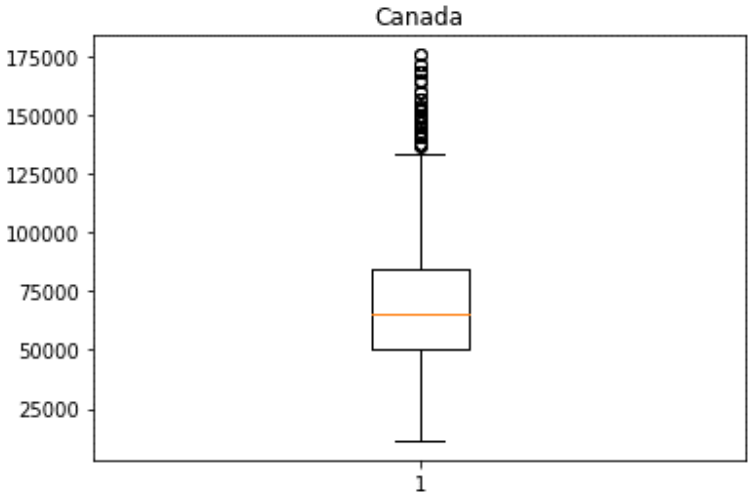
Para la variable **Country = Germany**:

count	2791.000000
mean	63727.367252
std	27054.126941
min	11352.000000
25%	48120.000000
50%	61870.000000
75%	76992.000000
max	171862.000000



Para la variable **Country = Canada**:

count	1697.000000
mean	70981.443724
std	29220.350447
min	11451.000000
25%	50383.000000
50%	65651.000000
75%	84736.000000
max	175579.000000



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:

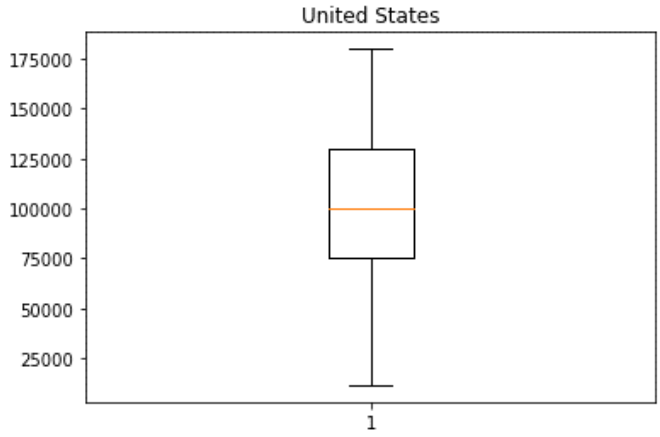
Country = United States

count	10589.000000
mean	101867.868543
std	36521.414682
min	11400.000000
25%	75000.000000
50%	100000.000000
75%	130000.000000
max	180000.000000

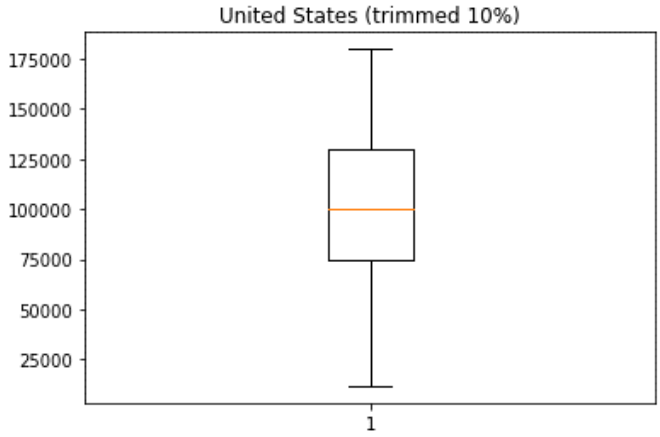
Datos originales

count	8473.000000
mean	101721.132067
std	36564.792119
min	11750.000000
25%	75000.000000
50%	100000.000000
75%	130000.000000
max	180000.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

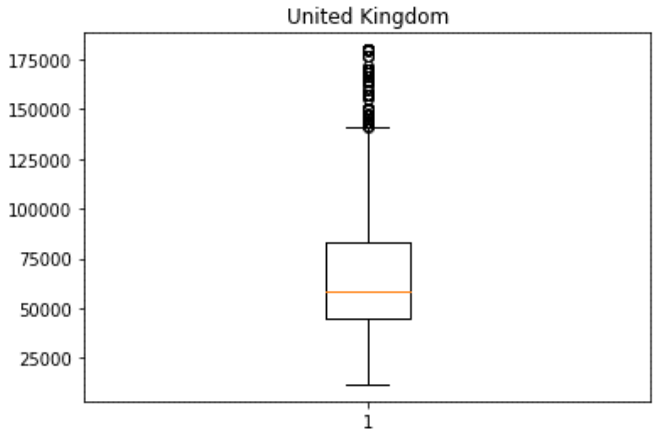
Country = United Kingdom

count	2887.000000
mean	67599.931417
std	32468.041138
min	11776.000000
25%	44488.000000
50%	58881.000000
75%	83218.500000
max	179915.000000

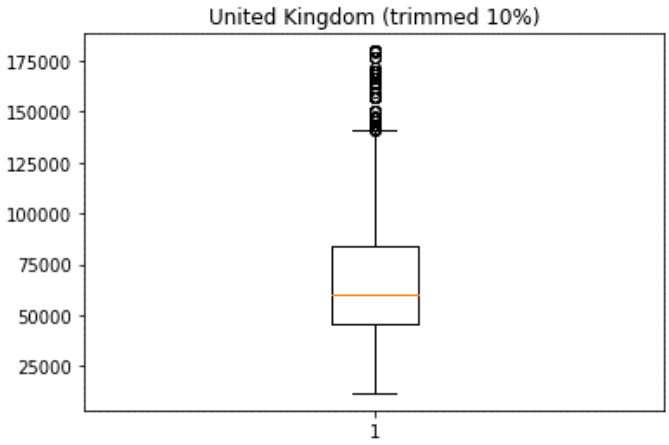
Datos originales

count	2311.000000
mean	68051.830376
std	32628.495998
min	11776.000000
25%	45797.000000
50%	60190.000000
75%	83742.000000
max	179915.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

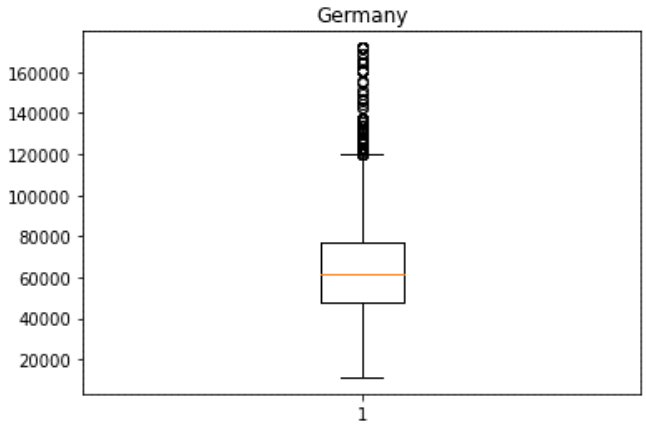
Country = Germany

count	2791.000000
mean	63727.367252
std	27054.126941
min	11352.000000
25%	48120.000000
50%	61870.000000
75%	76992.000000
max	171862.000000

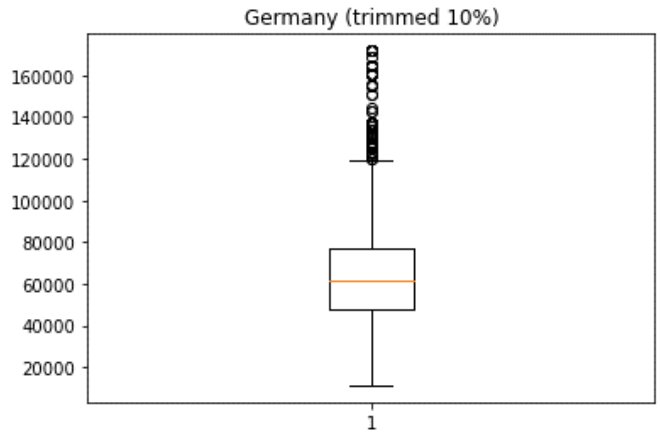
Datos originales

count	2233.000000
mean	63598.520824
std	26983.108078
min	11352.000000
25%	48120.000000
50%	61872.000000
75%	76765.000000
max	171862.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

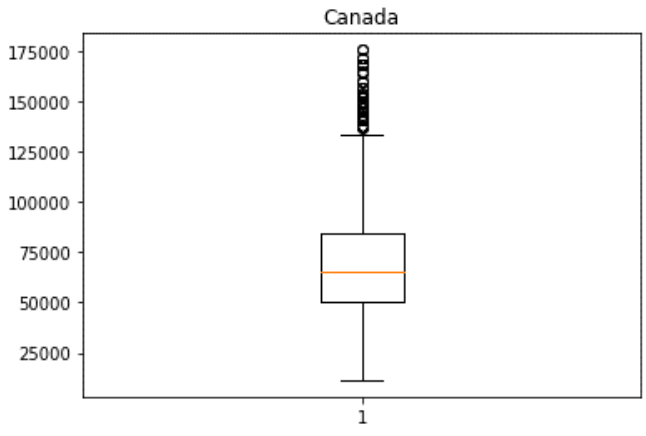
Country = Canada

count	1697.000000
mean	70981.443724
std	29220.350447
min	11451.000000
25%	50383.000000
50%	65651.000000
75%	84736.000000
max	175579.000000

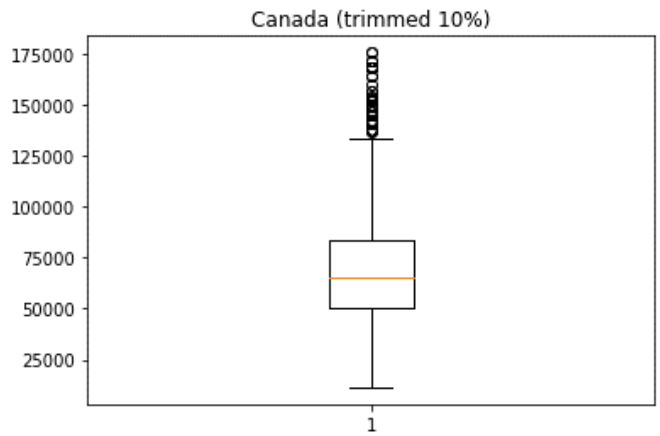
Datos originales

count	1359.000000
mean	70781.242090
std	29091.321077
min	11451.000000
25%	50383.000000
50%	65424.000000
75%	83972.000000
max	175579.000000

Datos recortados 10%



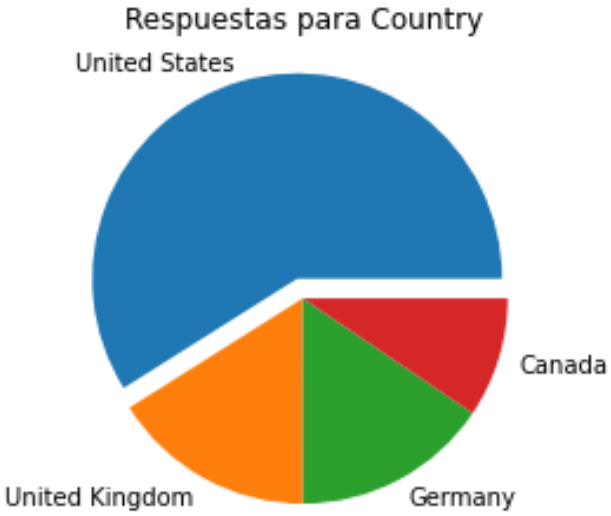
Datos originales



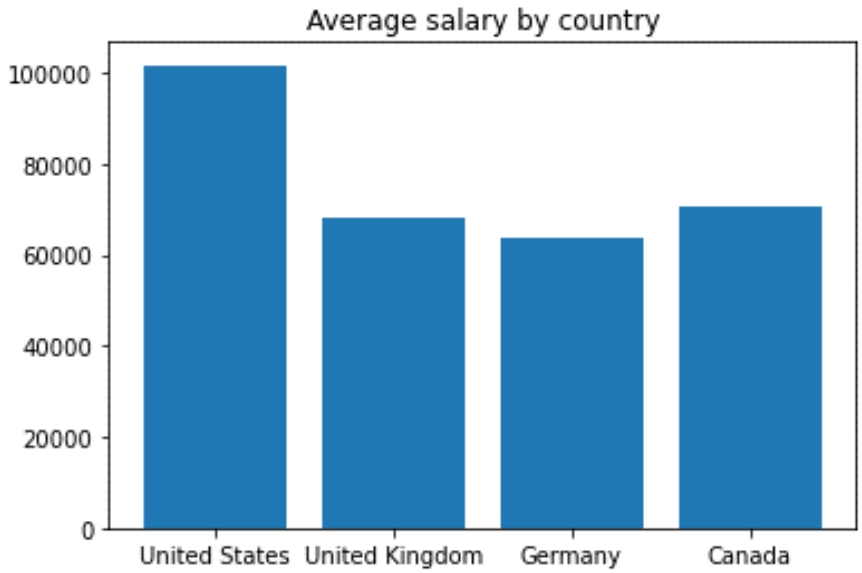
Datos recortados 10%

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.

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

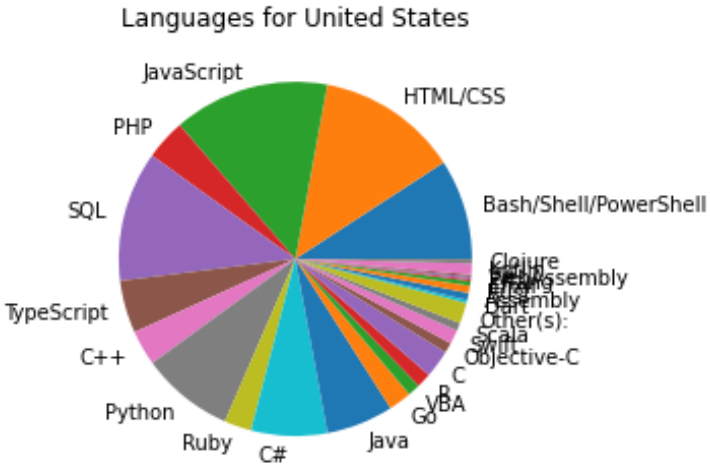


b. Haciendo un análisis de los salarios anuales promedio con respecto a los 4 países más comunes, se puede identificar que el país que tiende a tener los salarios más altos es ‘United States’ con un salario promedio de $101,721.13 y el país que tiende a tener los salarios más bajos es ‘Germany’ con un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



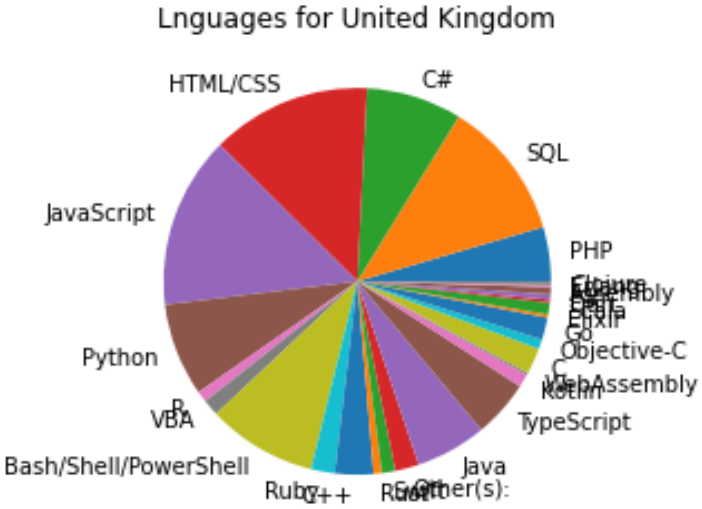
c. El análisis de los datos también nos muestra que los lenguajes de programación más populares y menos populares para cada uno de los 4 países más comunes son los que se muestran a continuación:

Country = United States



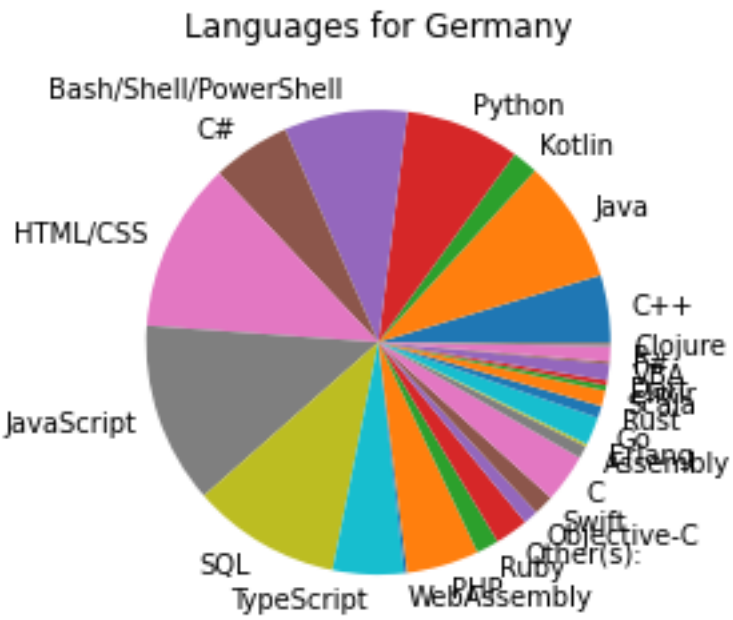
El lenguaje más popular es: JavaScript.
El lenguaje menos popular es: Erlang

Country = United Kingdom



El lenguaje más popular es: JavaScript
El lenguaje menos popular es: Erlang

Country = Germany



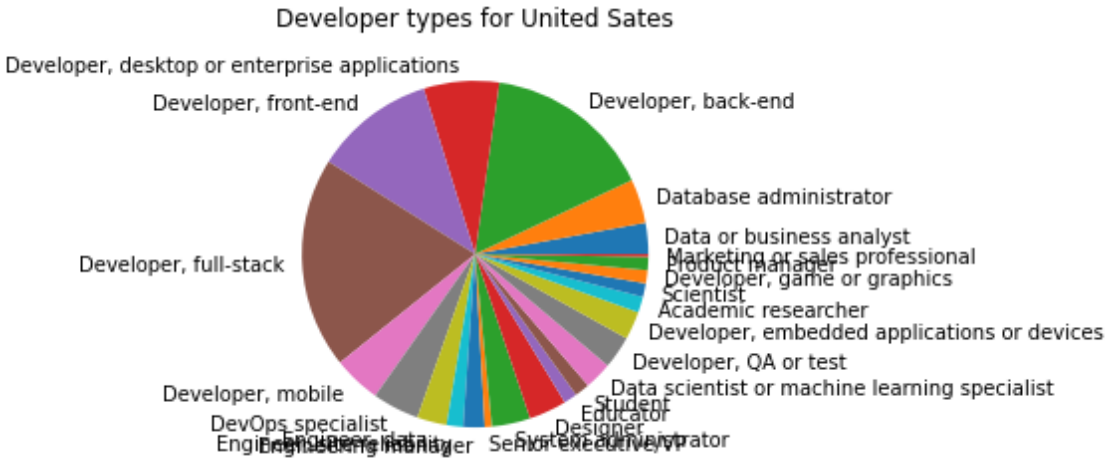
El lenguaje más popular es: JavaScript
El lenguaje menos popular es: F#

Country = Canada

El lenguaje más popular es: JavaScript
El lenguaje menos popular es: Erlang

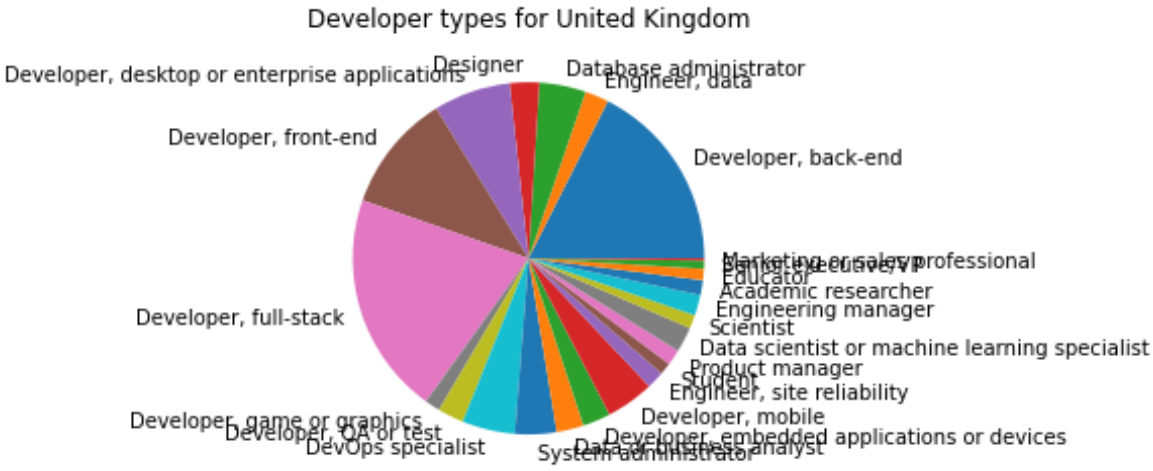
d. Al analizar los datos de la variable **DevType,** podemos encontrar cuales son los tipos de desarrollador más comunes para cada uno de los países más comunes:

Country = United States



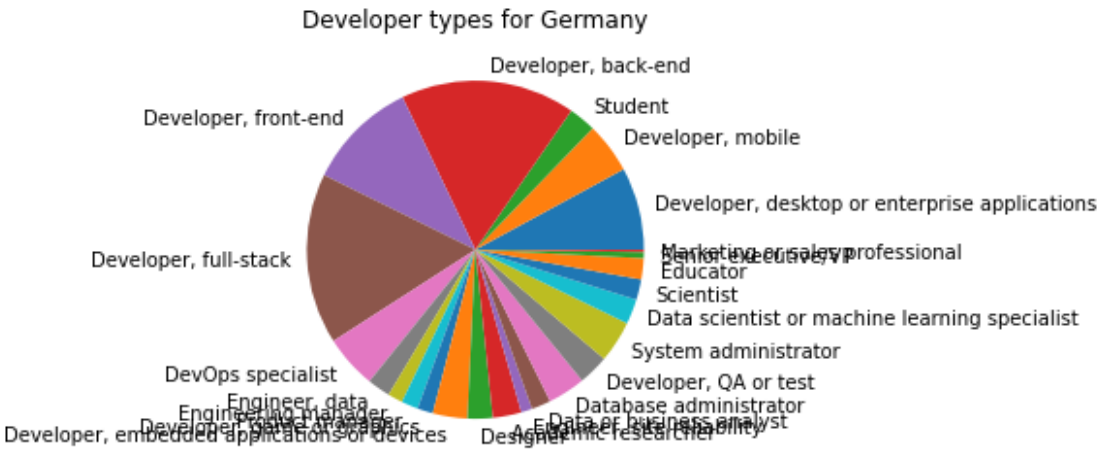
El tipo de desarrollador más popular es: Developer, full-stack.
El tipo de desarrollador menos popular es: Marketing or sales professional

Country = United Kingdom



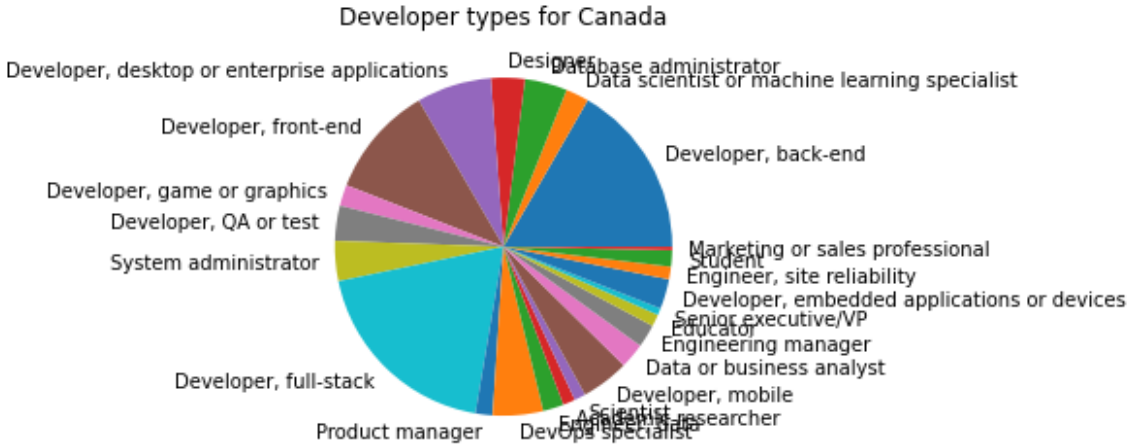
El tipo de desarrollador más popular es: Developer, full-stack
El tipo de desarrollador menos popular es: Marketing or sales professional

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

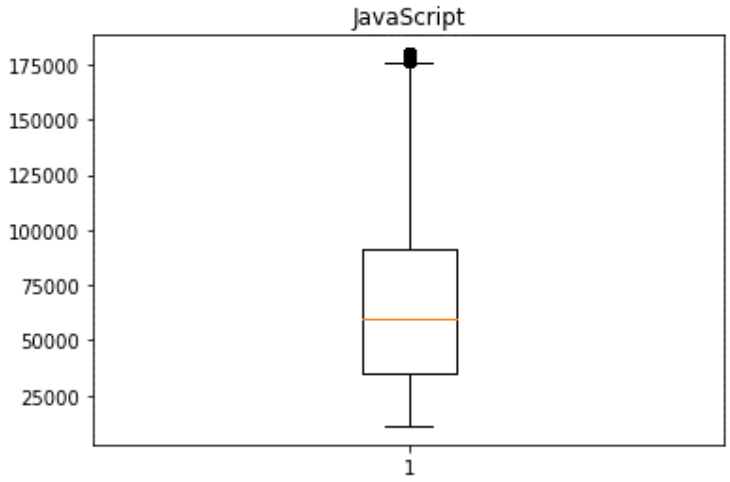


El tipo de desarrollador más popular es: Developer, full-stack
El tipo de desarrollador menos popular es: Marketing or sales professional

4. 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:

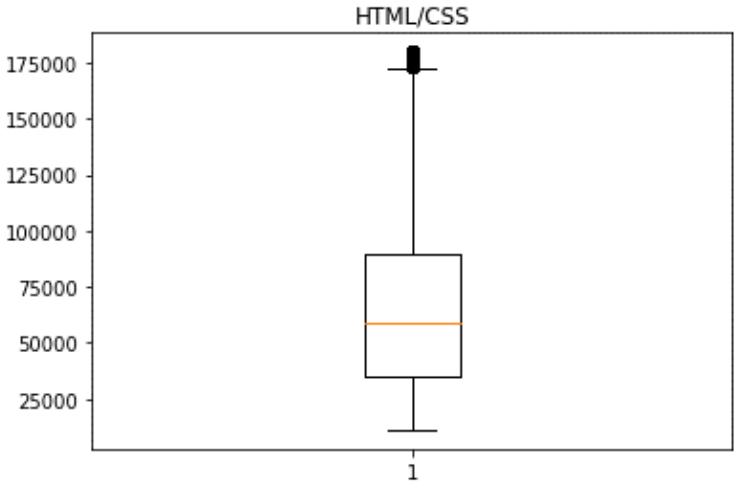
   Para la variable **LanguageWorkedWith = JavaScript:**

count	25804.000000
mean	66949.091381
std	39822.429085
min	11220.000000
25%	35000.000000
50%	59652.000000
75%	91593.000000
max	180000.000000



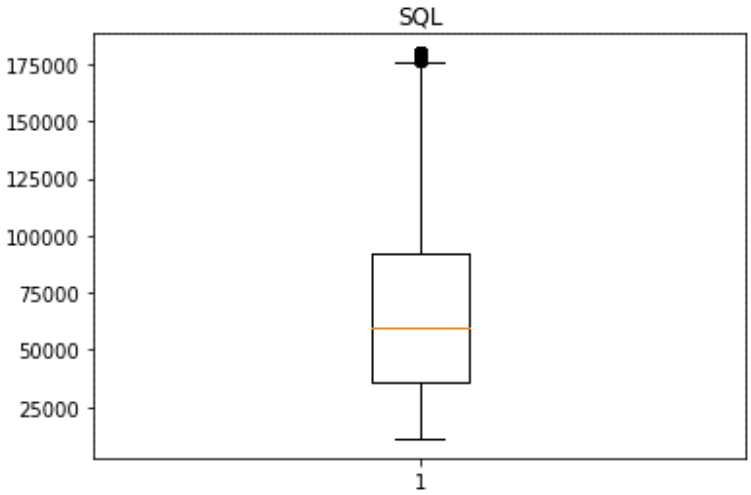
Para la variable **LanguageWorkedWith = HTML/CSS**:

count	23109.000000
mean	66229.947813
std	39355.412993
min	11220.000000
25%	34726.000000
50%	58881.000000
75%	90000.000000
max	180000.000000



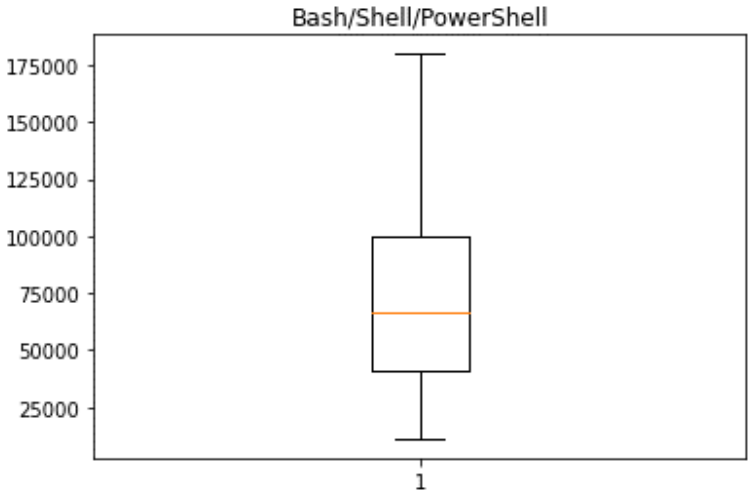
Para la variable **LanguageWorkedWith = SQL**:

count	21102.000000
mean	67534.560658
std	39783.876263
min	11220.000000
25%	35748.000000
50%	60000.000000
75%	92000.000000
max	180000.000000



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

count	15216.000000
mean	73222.353247
std	40561.835532
min	11220.000000
25%	41244.000000
50%	66453.000000
75%	100000.000000
max	180000.000000



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:

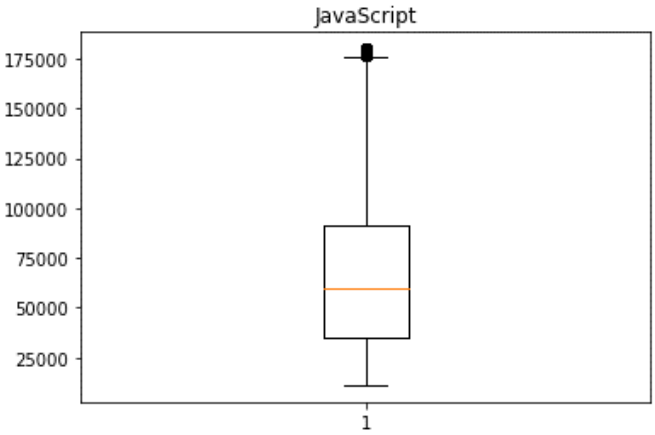
LanguageWorkedWith = JavaScript

count	25804.000000
mean	66949.091381
std	39822.429085
min	11220.000000
25%	35000.000000
50%	59652.000000
75%	91593.000000
max	180000.000000

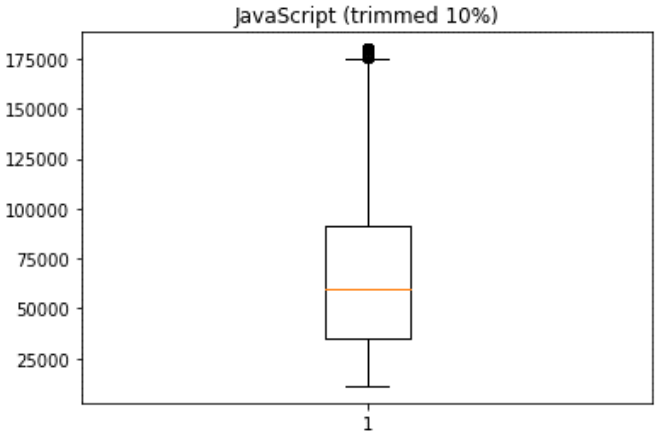
Datos originales

count	20644.000000
mean	66848.913098
std	39851.883117
min	11220.000000
25%	34872.000000
50%	59579.000000
75%	91130.250000
max	180000.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

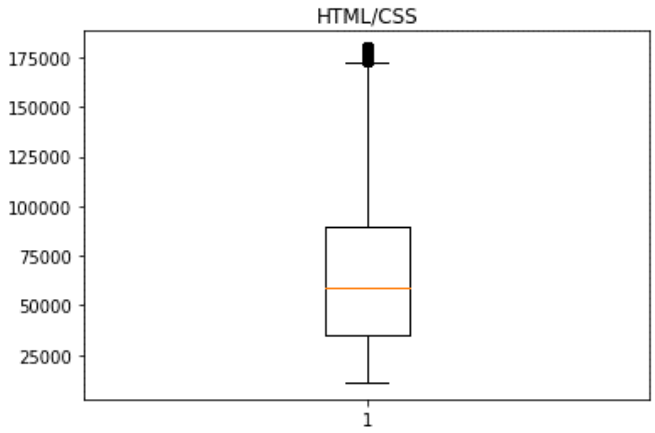
LanguageWorkedWith = HTML/CSS

count	23109.000000
mean	66229.947813
std	39355.412993
min	11220.000000
25%	34726.000000
50%	58881.000000
75%	90000.000000
max	180000.000000

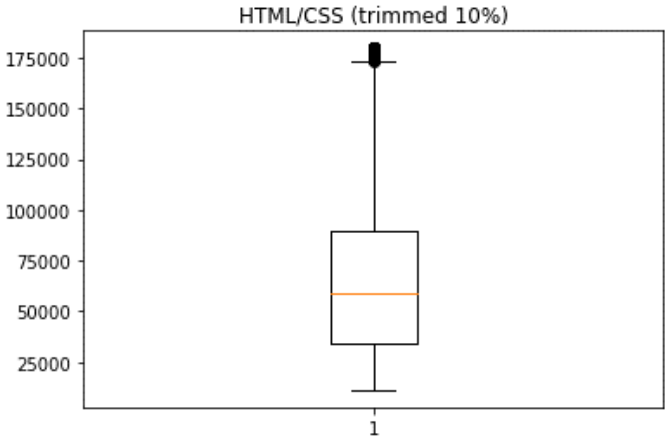
Datos originales

count	18489.000000
mean	66156.282871
std	39428.618685
min	11220.000000
25%	34500.000000
50%	58881.000000
75%	90000.000000
max	180000.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

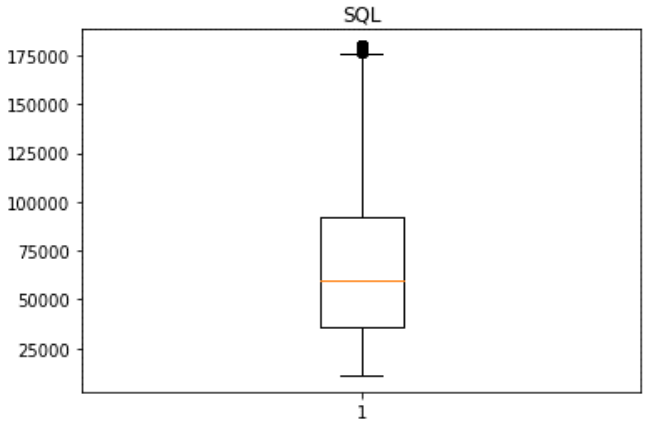
LanguageWorkedWith = SQL

count	21102.000000
mean	67534.560658
std	39783.876263
min	11220.000000
25%	35748.000000
50%	60000.000000
75%	92000.000000
max	180000.000000

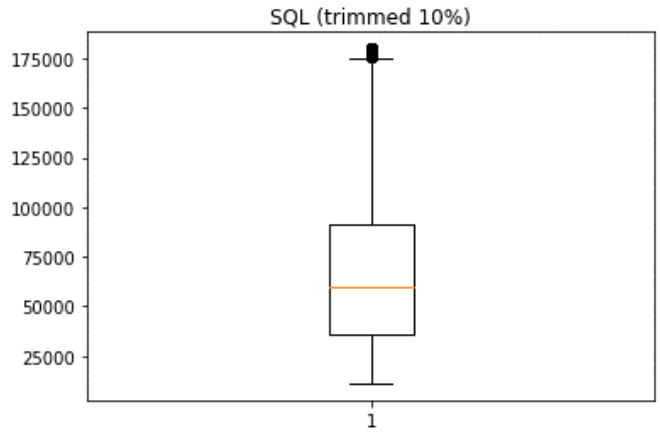
Datos originales

count	16882.000000
mean	67524.663843
std	39808.801649
min	11220.000000
25%	35748.000000
50%	60000.000000
75%	91752.000000
max	180000.000000

Datos recortados 10%



Datos originales



Datos recortados 10%

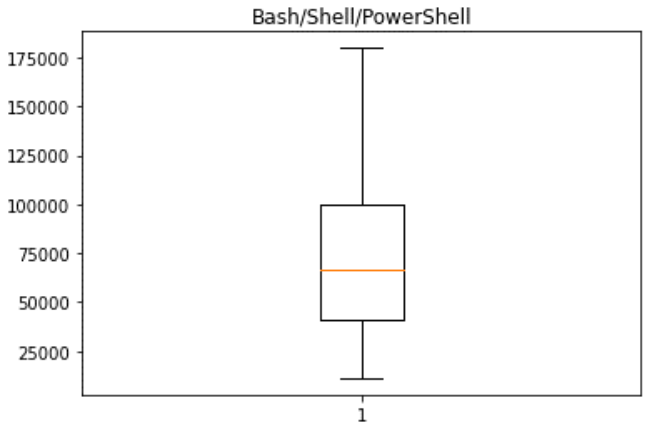
LanguageWorkedWith = Bash/Shell/PowerShell

count	15216.000000
mean	73222.353247
std	40561.835532
min	11220.000000
25%	41244.000000
50%	66453.000000
75%	100000.000000
max	180000.000000

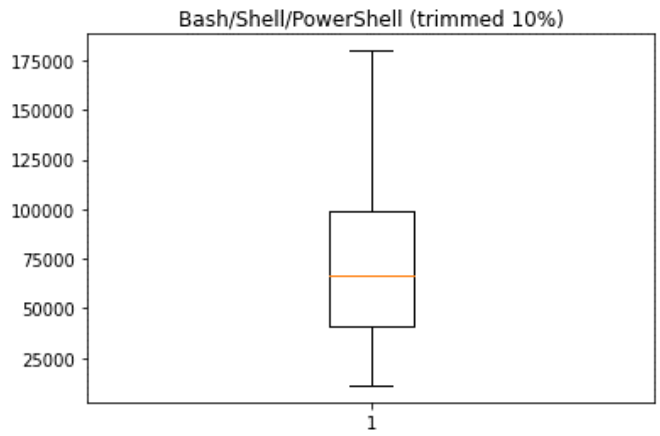
Datos originales

count	12174.000000
mean	73100.830951
std	40610.273578
min	11220.000000
25%	41244.000000
50%	66453.000000
75%	99266.250000
max	180000.000000

Datos recortados 10%



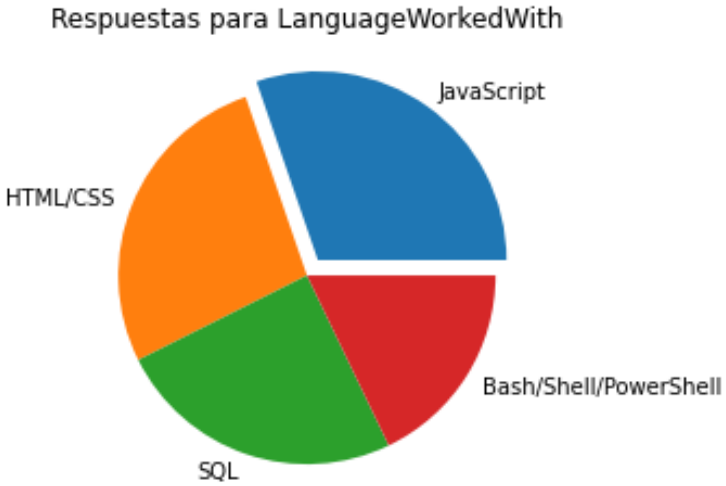
Datos originales



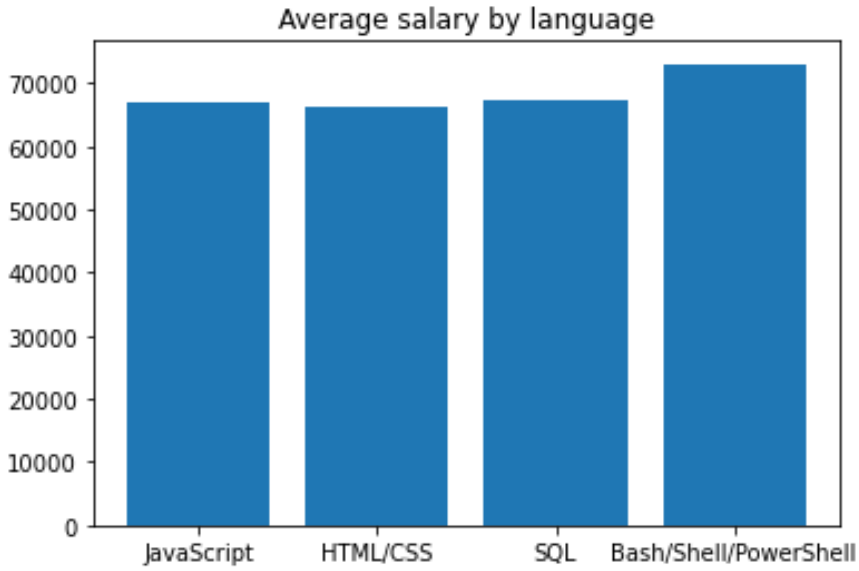
Datos recortados 10%

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.

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

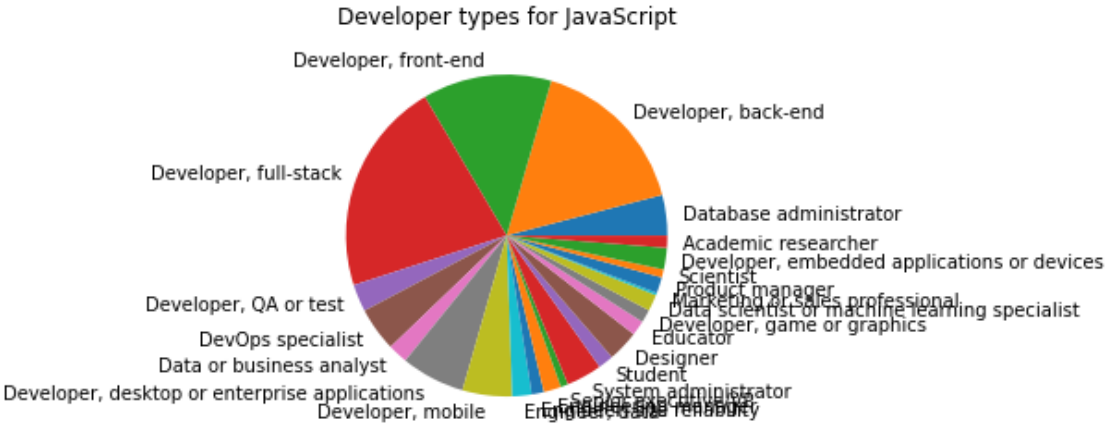


b. Haciendo un análisis de los salarios anuales promedio con respecto a los 4 lenguajes de programación más comunes, se puede identificar que el lenguaje que tiende a tener los salarios más altos es 'Bash/Shell/PowerShell' con un salario promedio de $73,100.83 y el lenguaje que tiende a tener los salarios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



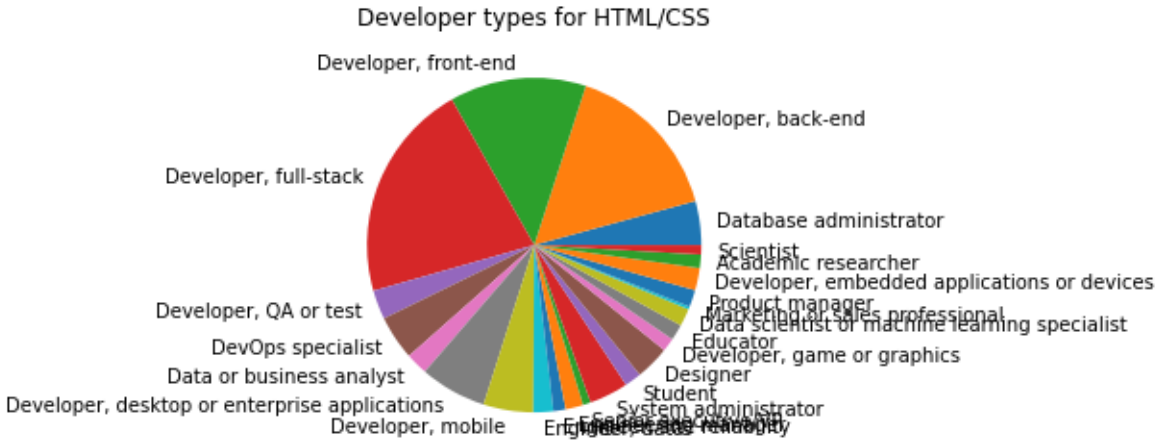
c. Al analizar los datos de la variable **DevType,** podemos encontrar cuales son los tipos de desarrollador más comunes para cada uno de los lenguajes de programación más comunes:

LanguageWorkedWith = JavaScript



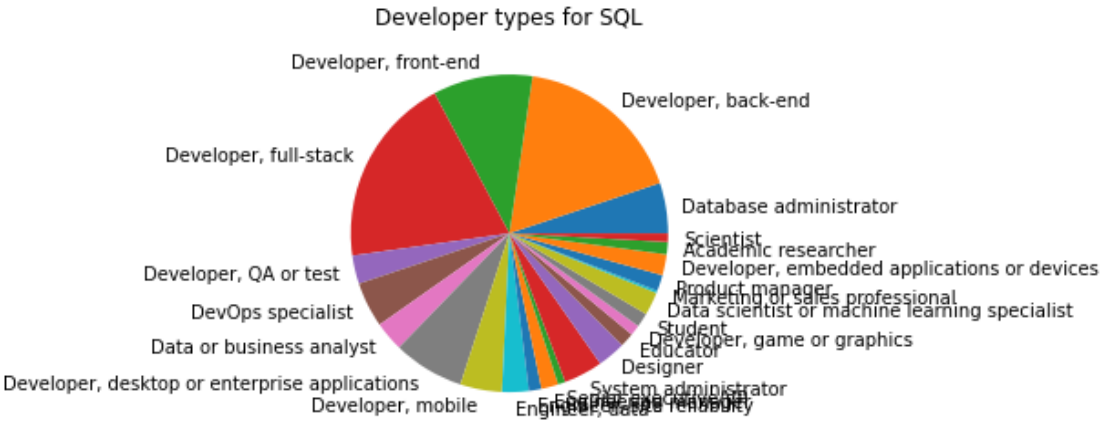
El tipo de desarrollador más popular es: Developer, full-stack.
El tipo de desarrollador menos popular es: Marketing or sales professional

LanguageWorkedWith = HTML/CSS



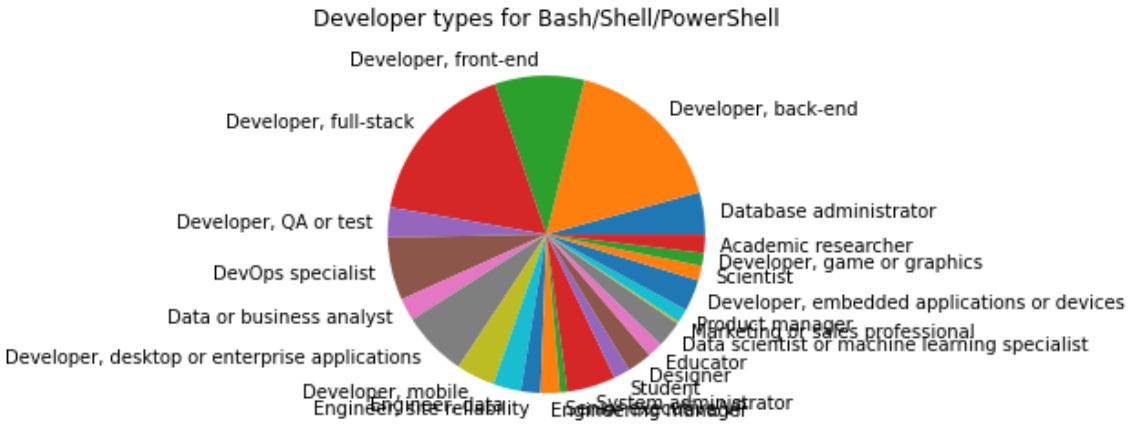
El tipo de desarrollador más popular es: Developer, full-stack
El tipo de desarrollador menos popular es: Marketing or sales professional

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