Minería de Datos

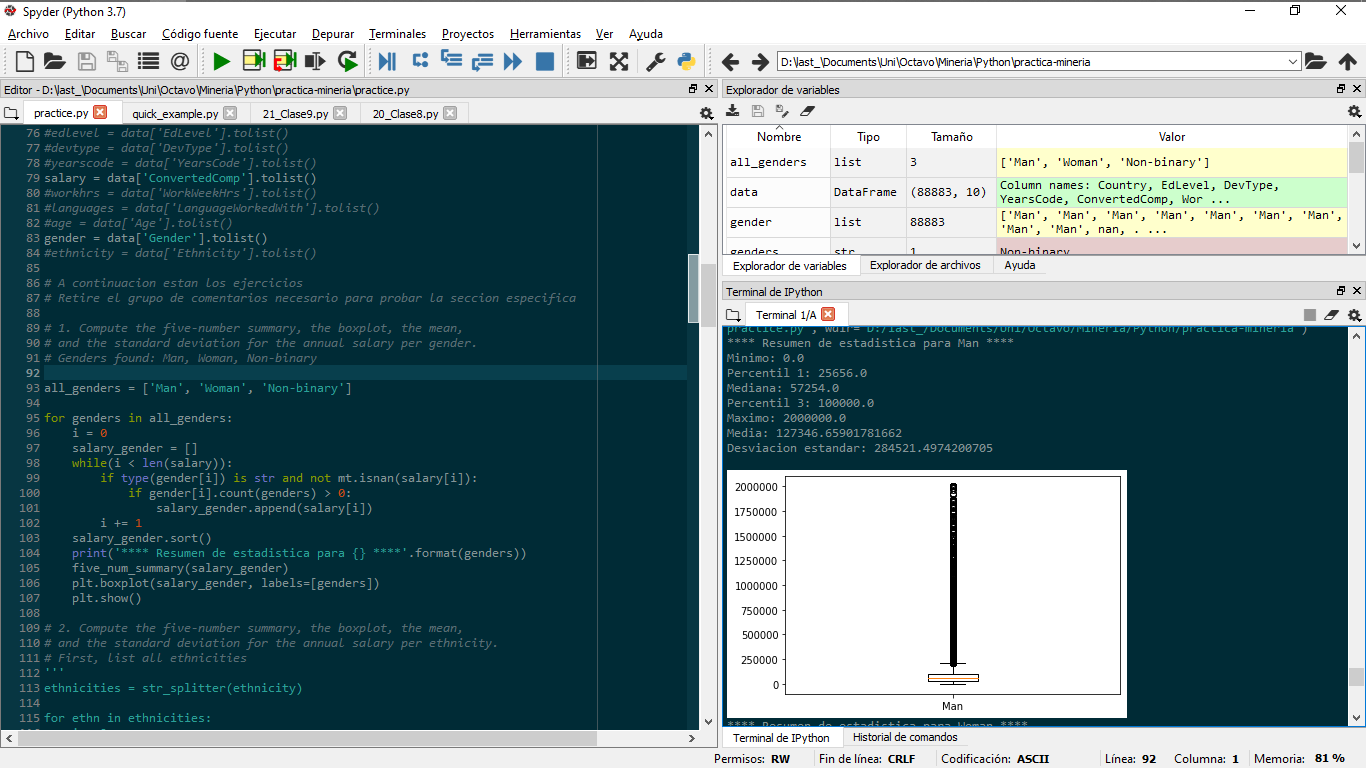
Práctica individual

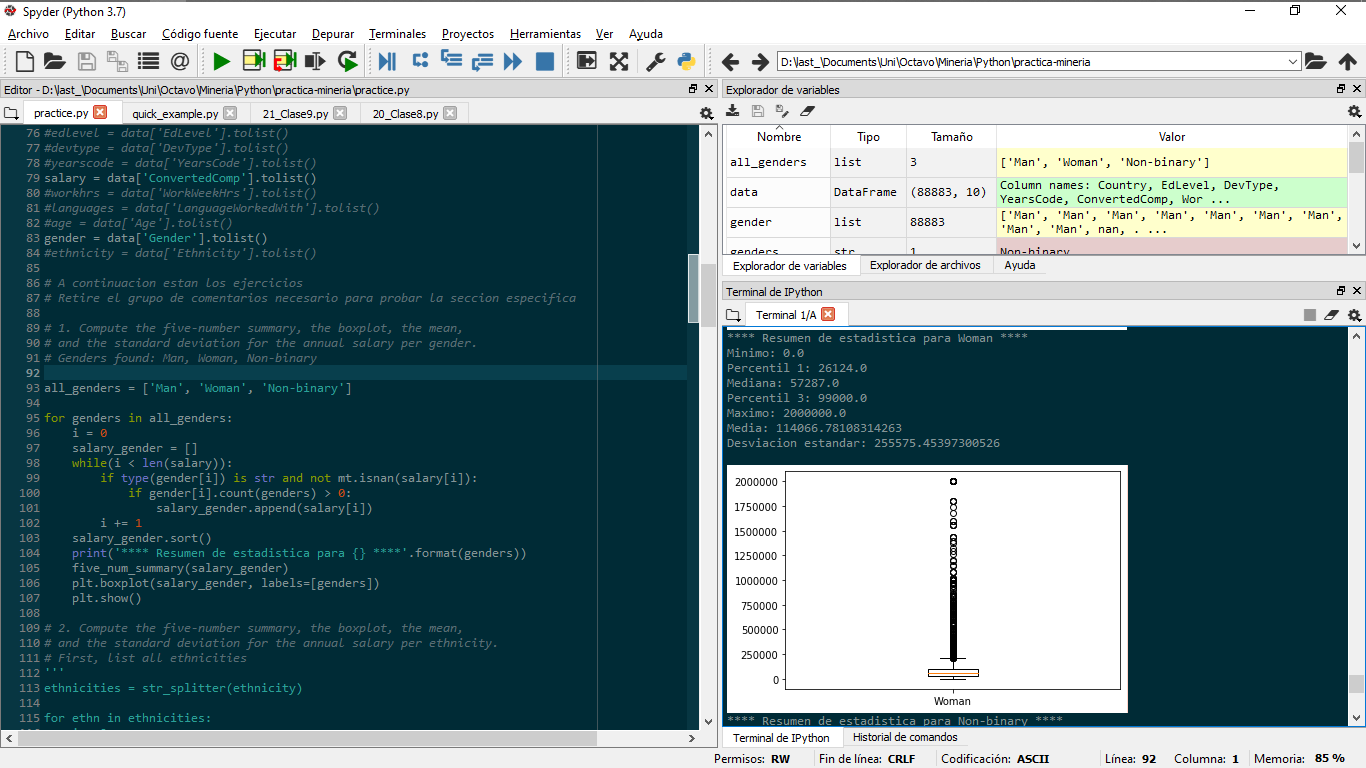
La práctica consiste en diversos programas cuyo propósito es analizar los resultados de una encuesta realizada a diversos usuarios de Stack OverFlow en 2019. Los datos recolectados van desde el país de residencia de los usuarios, su edad y el tiempo que han practicado programación hasta su salario anual o el puesto que practican en la sociedad.

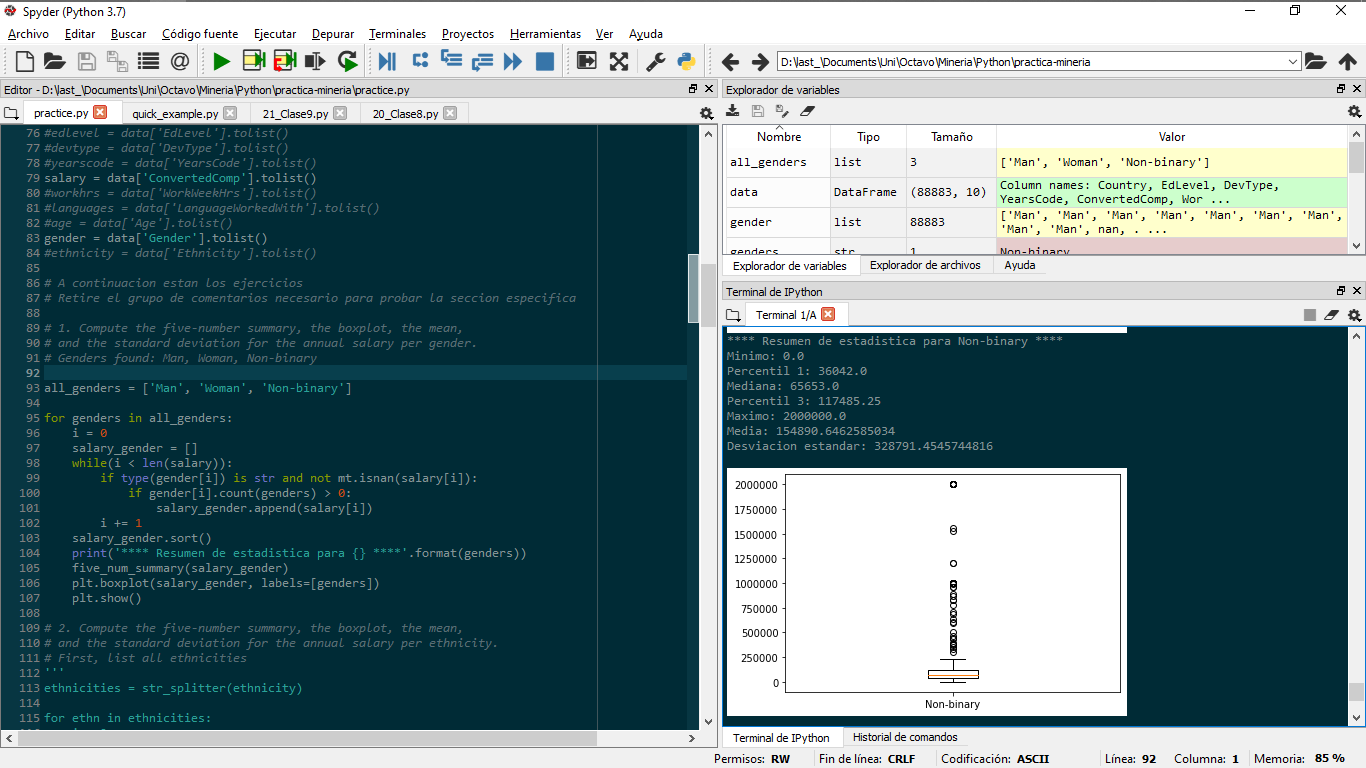
A continuación, se analizará punto por punto los diversos datos con tal de obtener información y conclusiones referentes a los mismos.

1. **Calcular el resumen de los 5 números, media, desviación estándar y boxplot del salario anual por género.**

Dentro de los resultados de la encuesta encontramos tres potenciales géneros: hombre, mujer y no binario. Con esto en mente, se procede a realizar el análisis de cada uno.



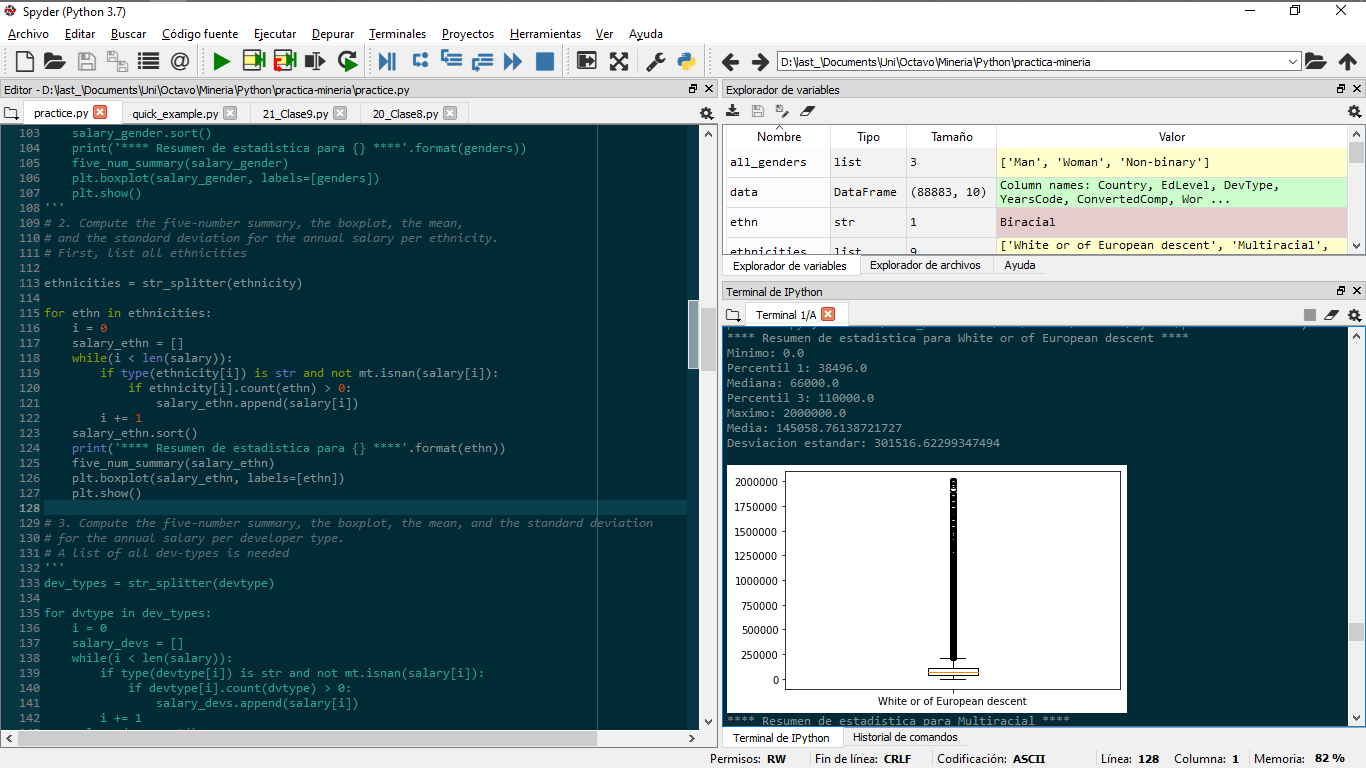


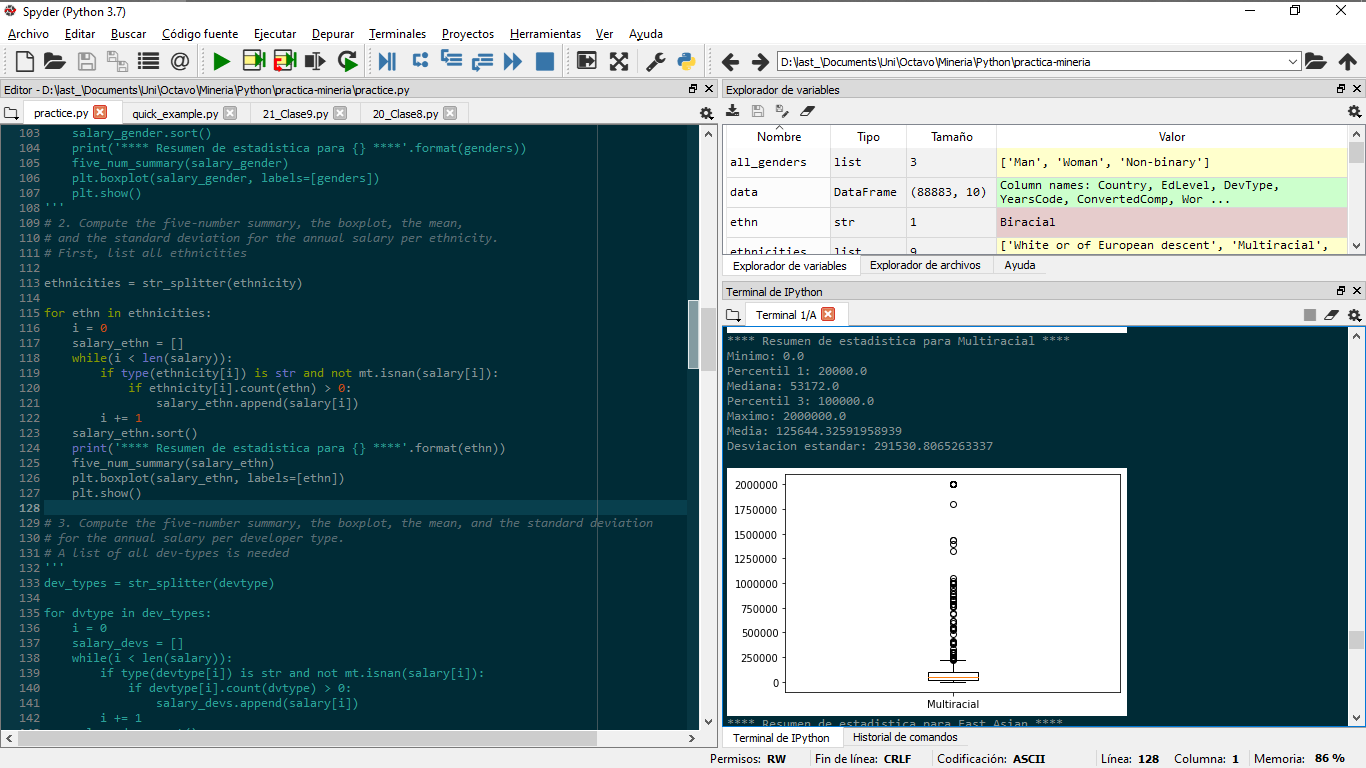
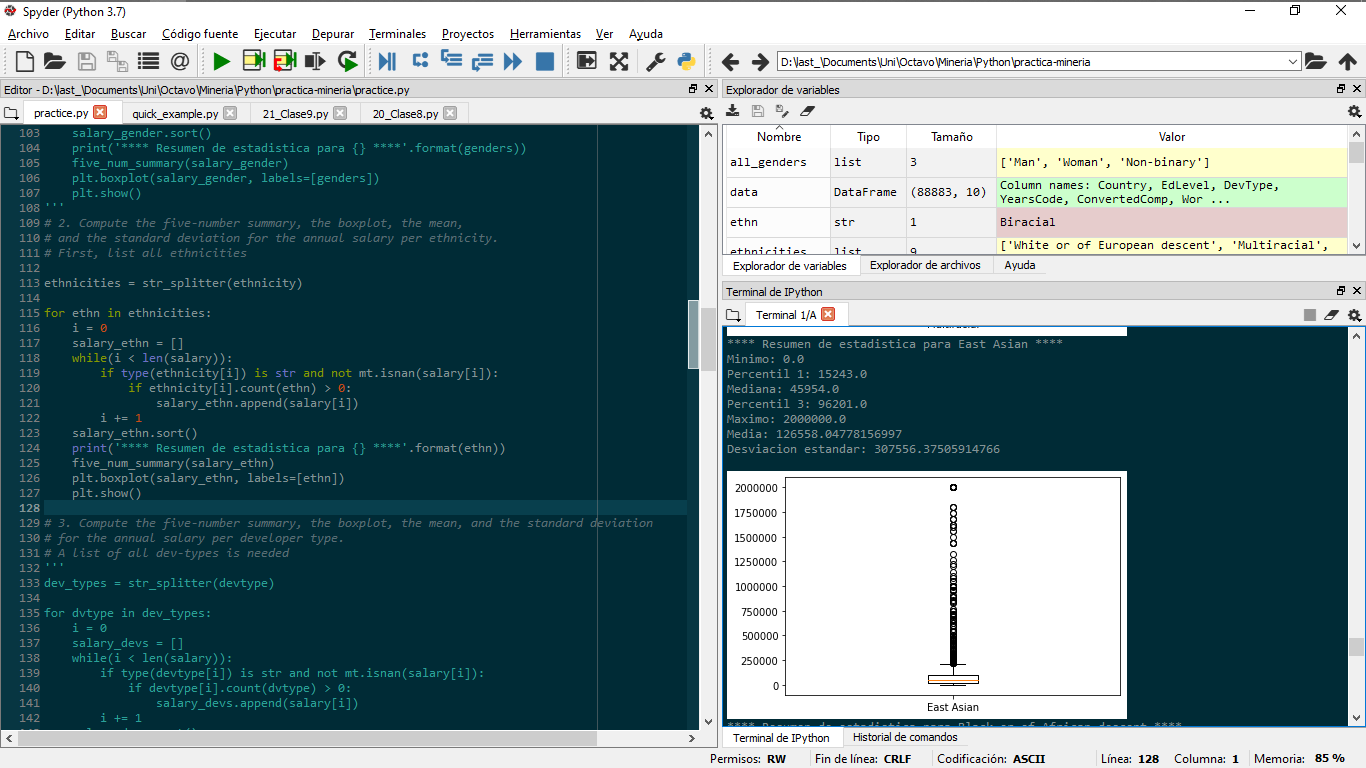


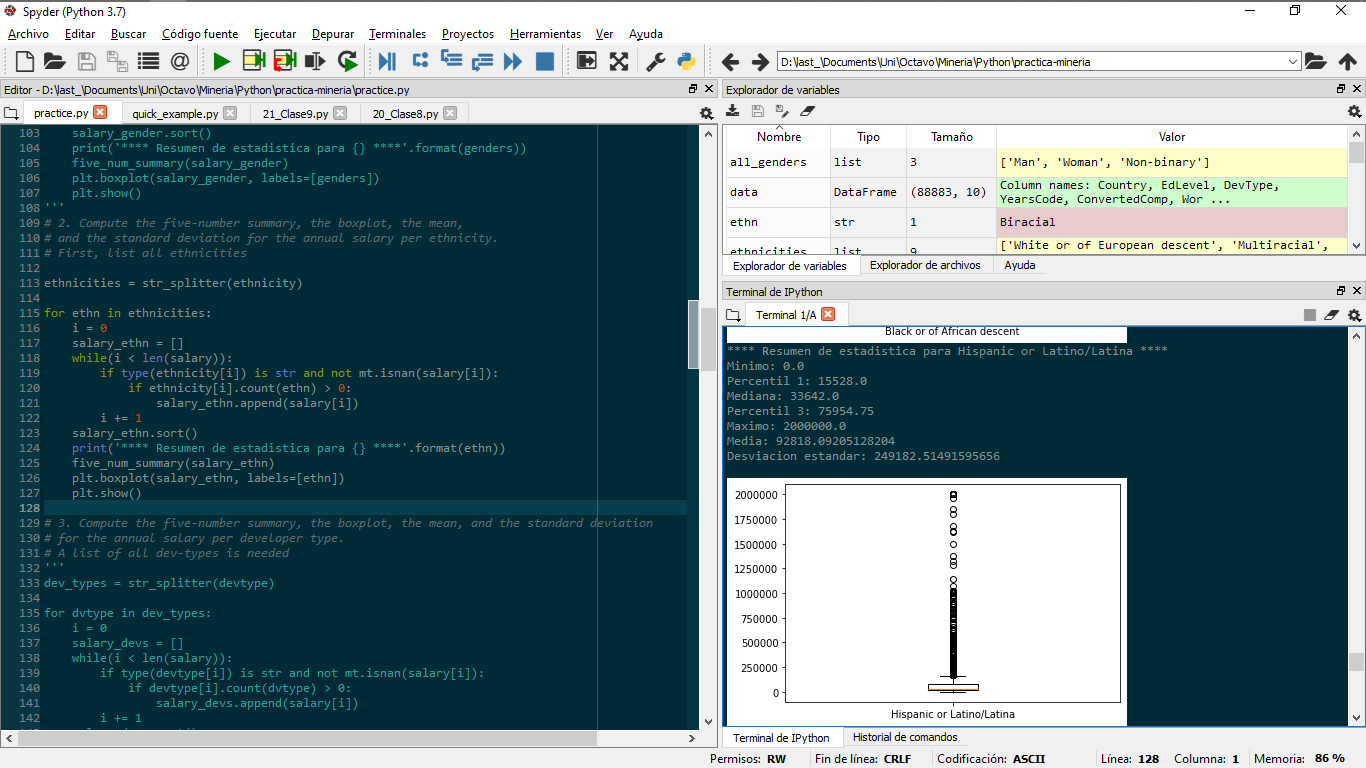
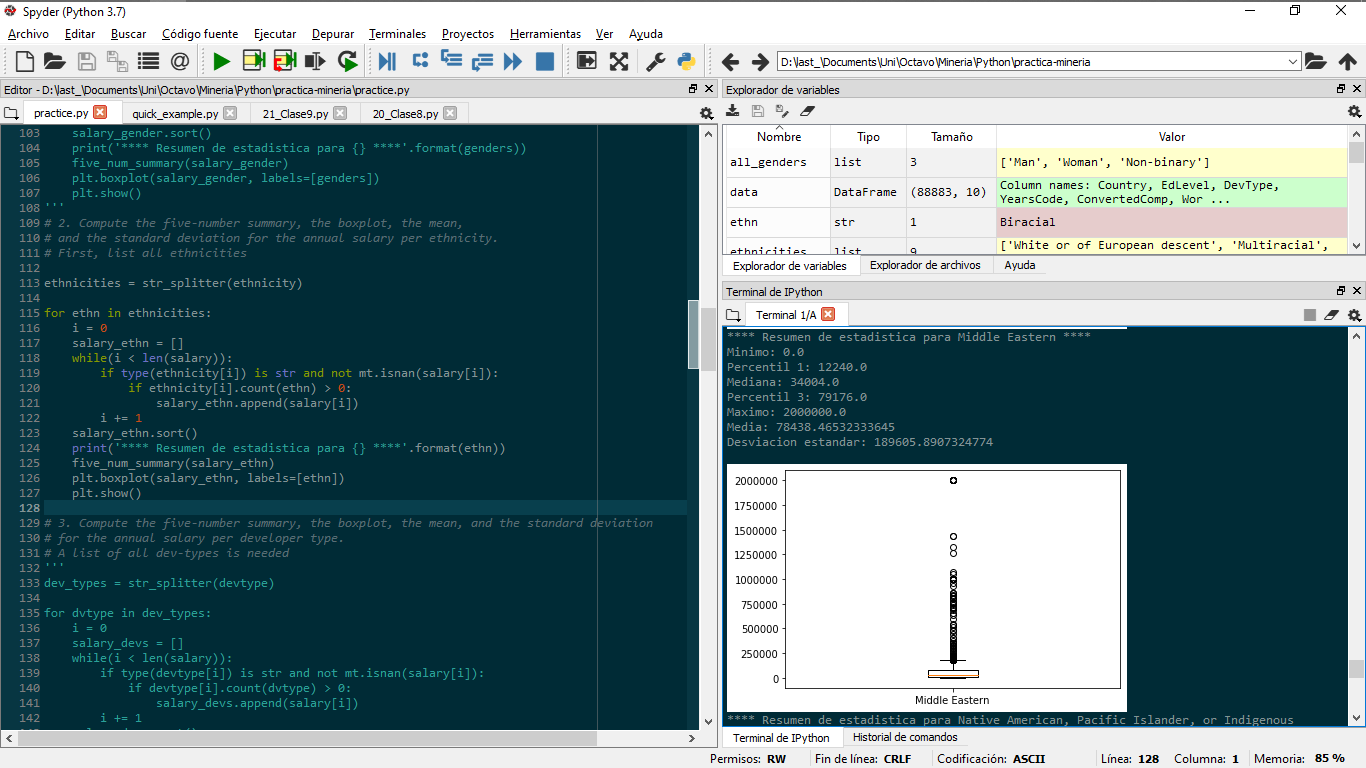
* La distribución indica que, aunque existan personas que ganan un máximo de $2M de dólares, la mayoría de las personas tienen un saldo equivalente o inferior a $200k dólares aproximadamente, independiente del género.
* A pesar de existir una vasta cantidad de *outliers*, es importante notar que la mayoría de las personas tienen un sueldo en el rango que se puede apreciar en los boxplot.
* Los registros de personas no-binarias ocupan una minoría respecto a los datos registrados, mientras los hombres ocupan la mayoría. A pesar de esto, el salario medio de las personas no-binarias supera a la media de los hombres. Un dato bastante curioso, a mi parecer.
* Cabe destacar que se encontraron datos de personas trabajando gratis (con ganancias de 0). Un dato encontrado gracias a conocer el mínimo salario posible.

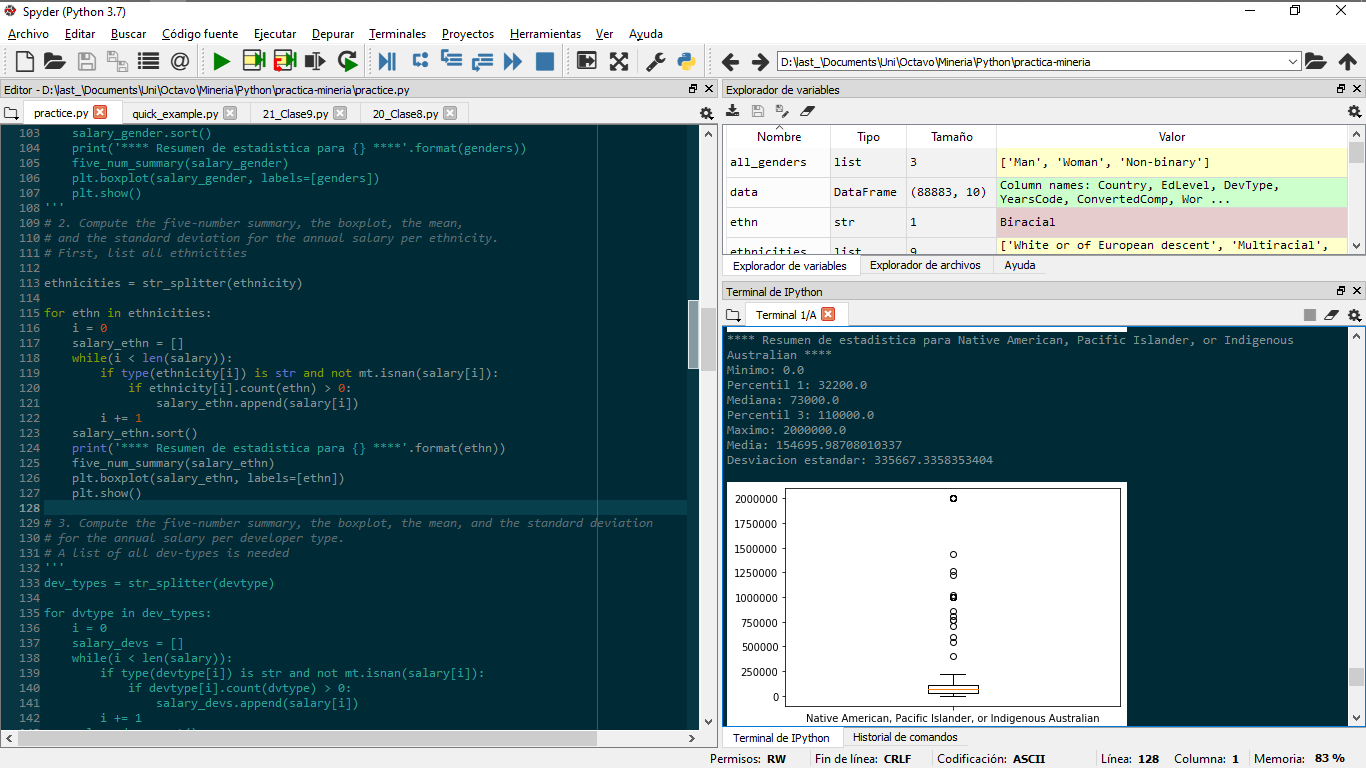
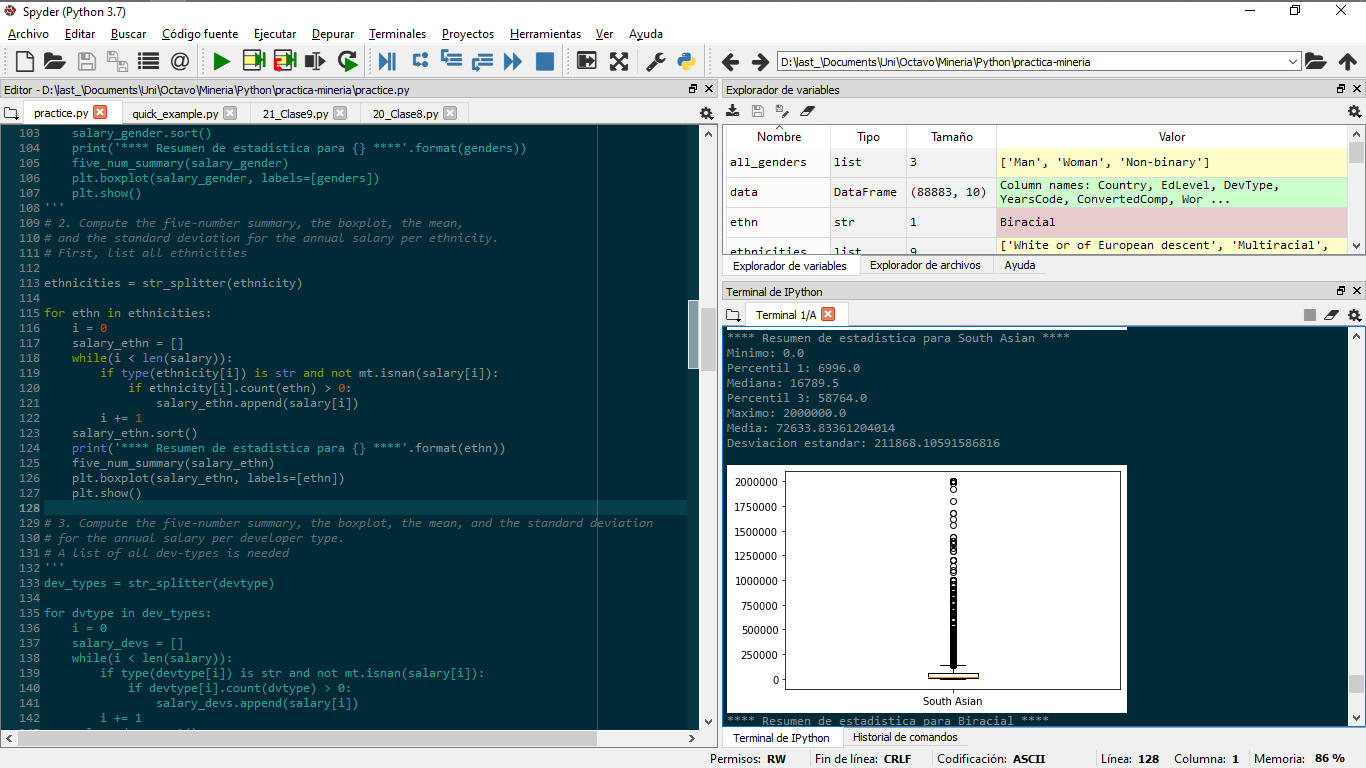
1. **Calcular el resumen de los 5 números, media, desviación estándar y boxplot del salario anual por etnia.**

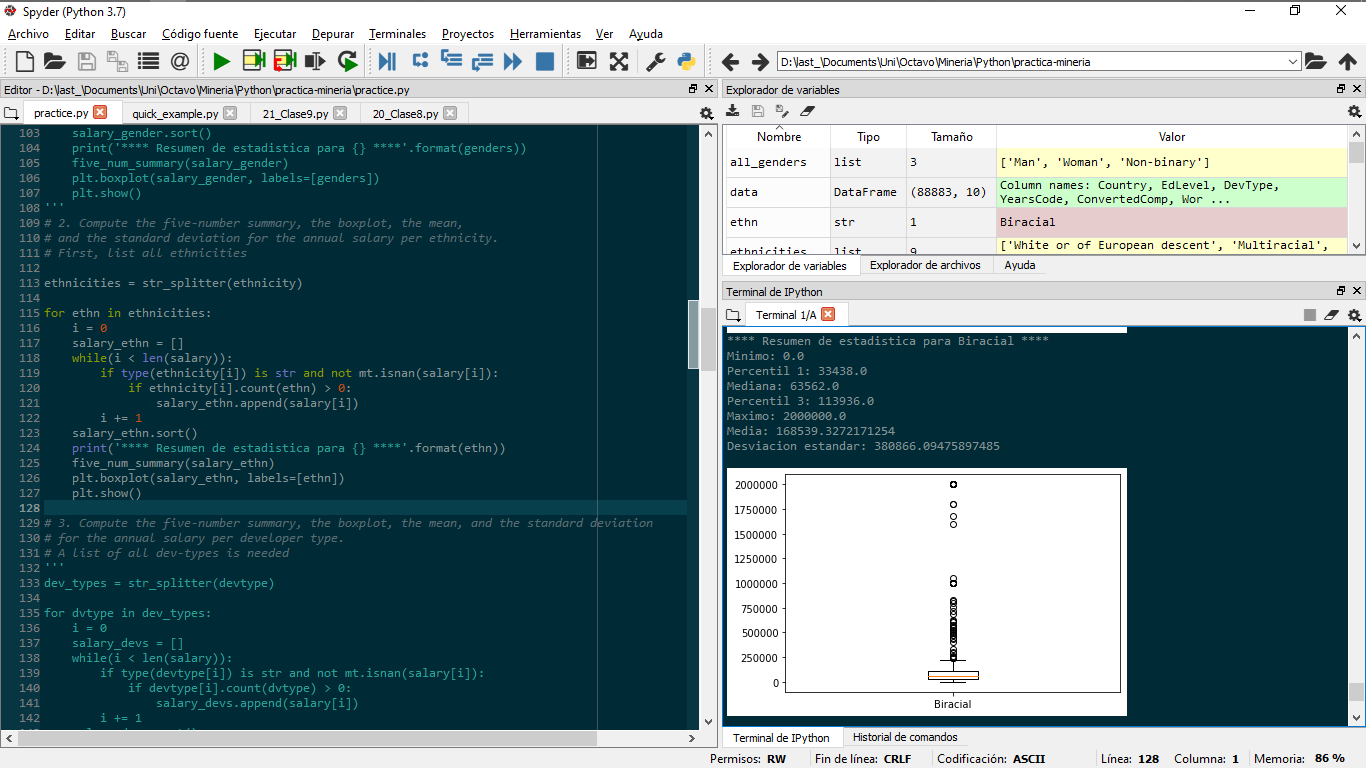
Se encuentran varias etnias base, junto a las opciones “biracial” y “multiracial”. A continuación, se exponen los resultados del análisis de cada una:



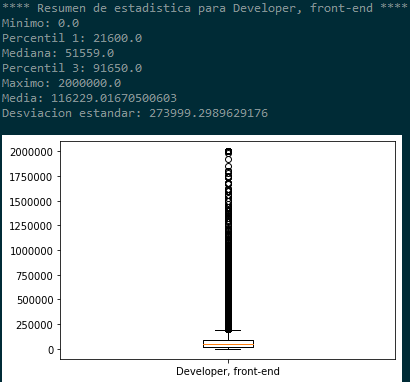
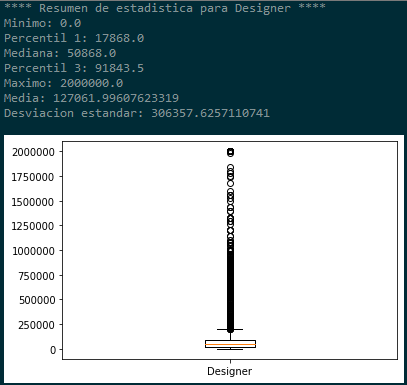


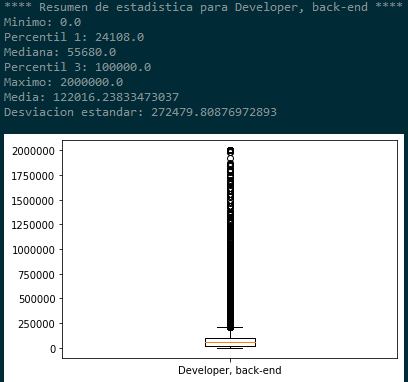
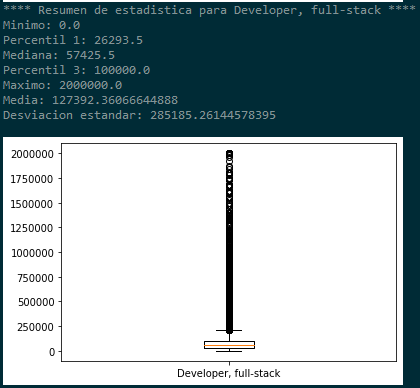
* Los resultados indican que la mayoría de las personas ganan el equivalente a $150k dólares o una cantidad menor. Esto concuerda con los datos de ganancias respecto al género.
* Una abrumadora cantidad de gente es caucásica, y su ganancia media ronda los $150k dólares.
* Mientras que las personas biraciales son quienes en promedio ganan más, las personas del sur de África no son tan afortunadas, pues la media de sus salarios anuales es por mucho inferior al resto de etnias. Esto también es bastante notorio en el respectivo boxplot.
* Aparentemente no hay muchos registros de nativos americanos respecto a otras etnias, pero no por eso su rango de ganancias es diferente respecto al resto.

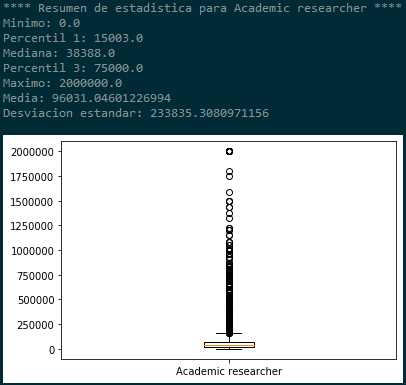
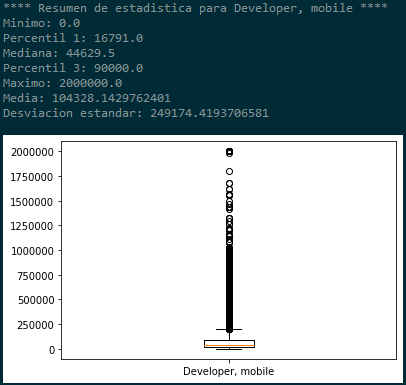
1. **Calcular el resumen de los 5 números, media, desviación estándar y boxplot del salario anual por tipo de desarrollador.**

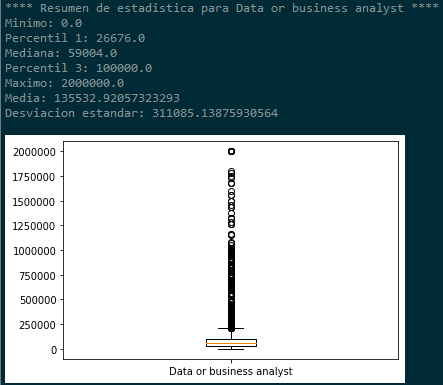
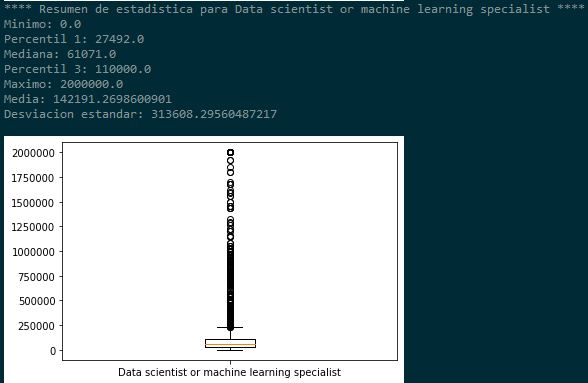
Existen diversos tipos de desarrollador dentro de las encuestas, desde estudiantes hasta científicos, o incluso ingenieros.

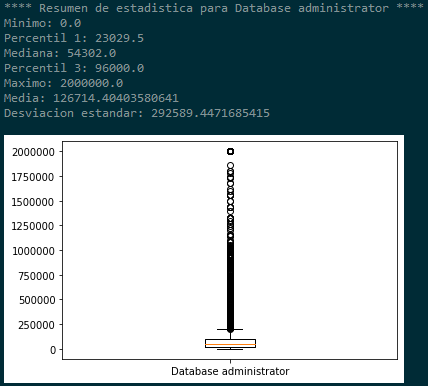
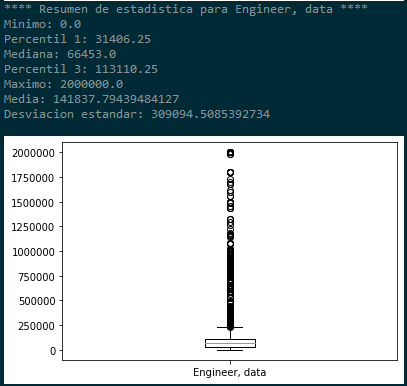


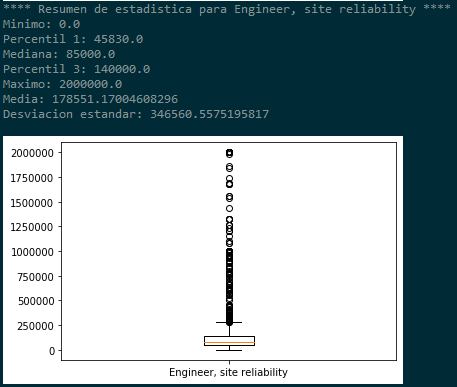
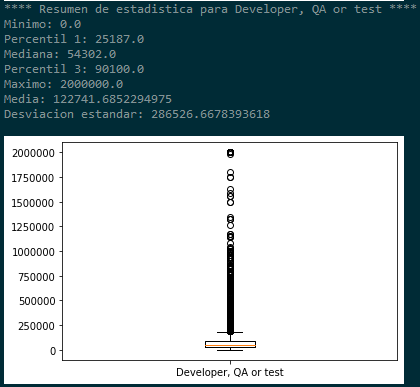
 

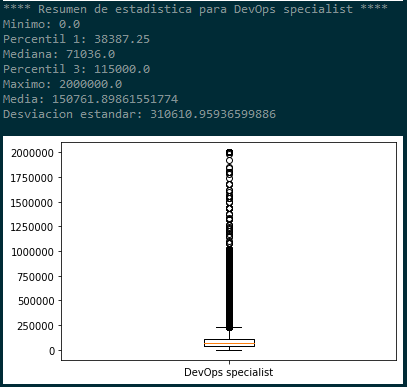
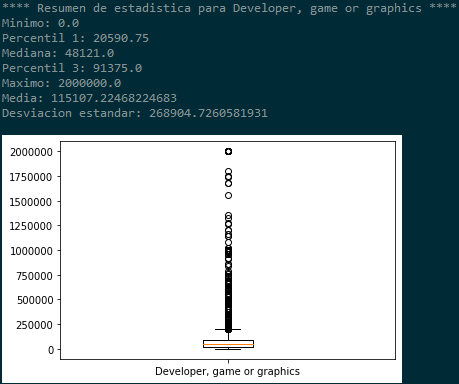
 

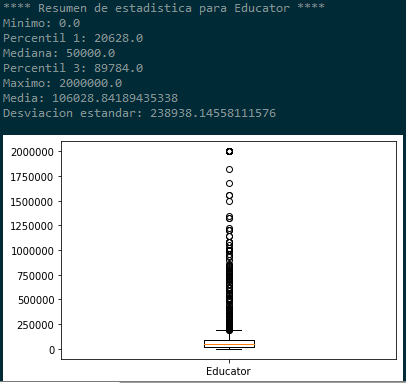
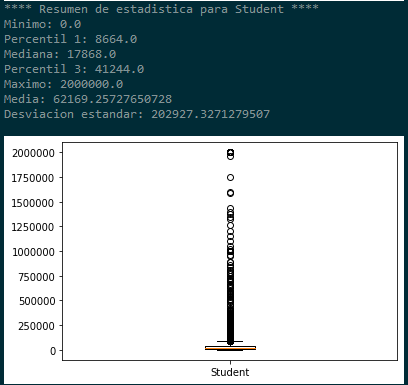
 

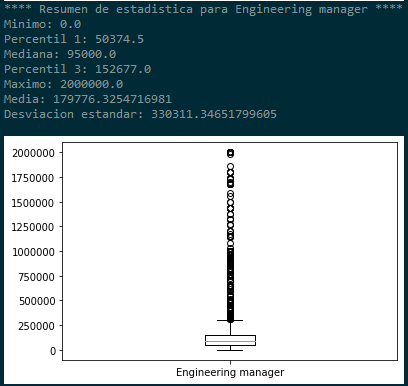
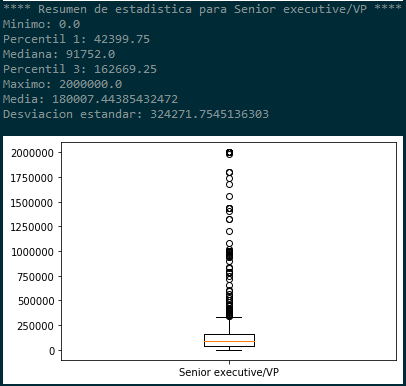
 

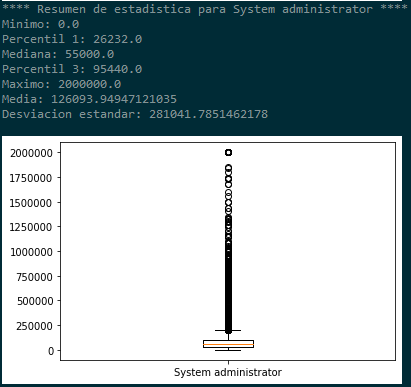
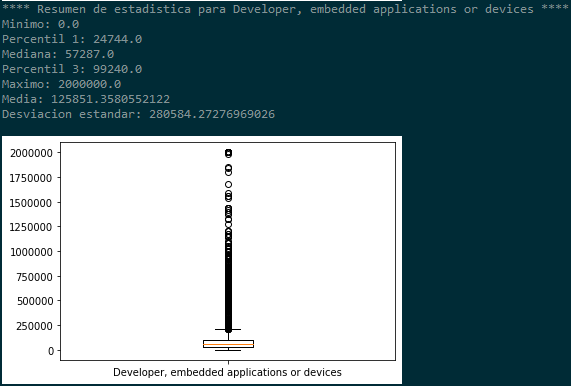
 

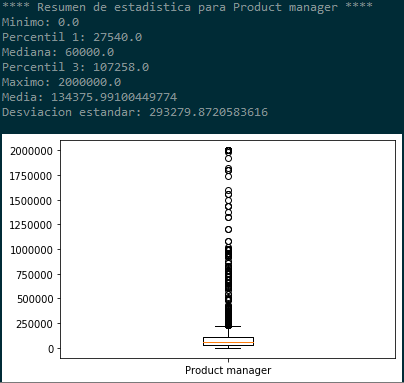
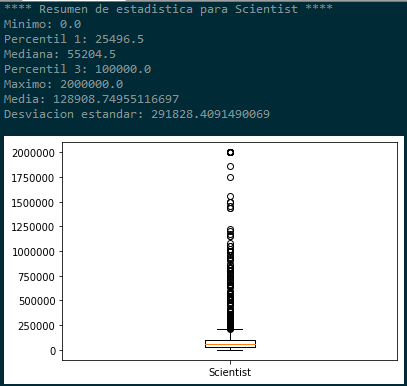
 

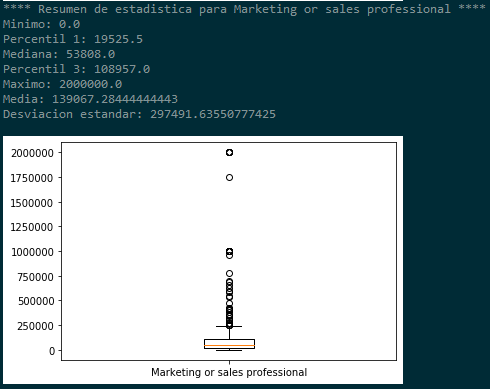
 



* Los rangos del salario son casi invariables respecto a los dos puntos anteriores, pues al parecer las personas siguen ubicándose en el mismo intervalo, independientemente de su puesto.
* Cabe mencionar que existen varias personas que, aunque tengan un puesto oficial de acuerdo con los resultados de la encuesta, poseen salario nulo (0).
* El grupo de personas con menor salario son los estudiantes, mientras que aquéllos con más ganancias son los ingenieros y los VP.
* Los estudiantes son los únicos cuyo boxplot se ve notablemente alterado con respecto al resto, aunque esto se debe a que su salario es mucho menor que el resto de las posiciones.

1. **Calcular la media, mediana y desviación estándar del salario anual por país.**

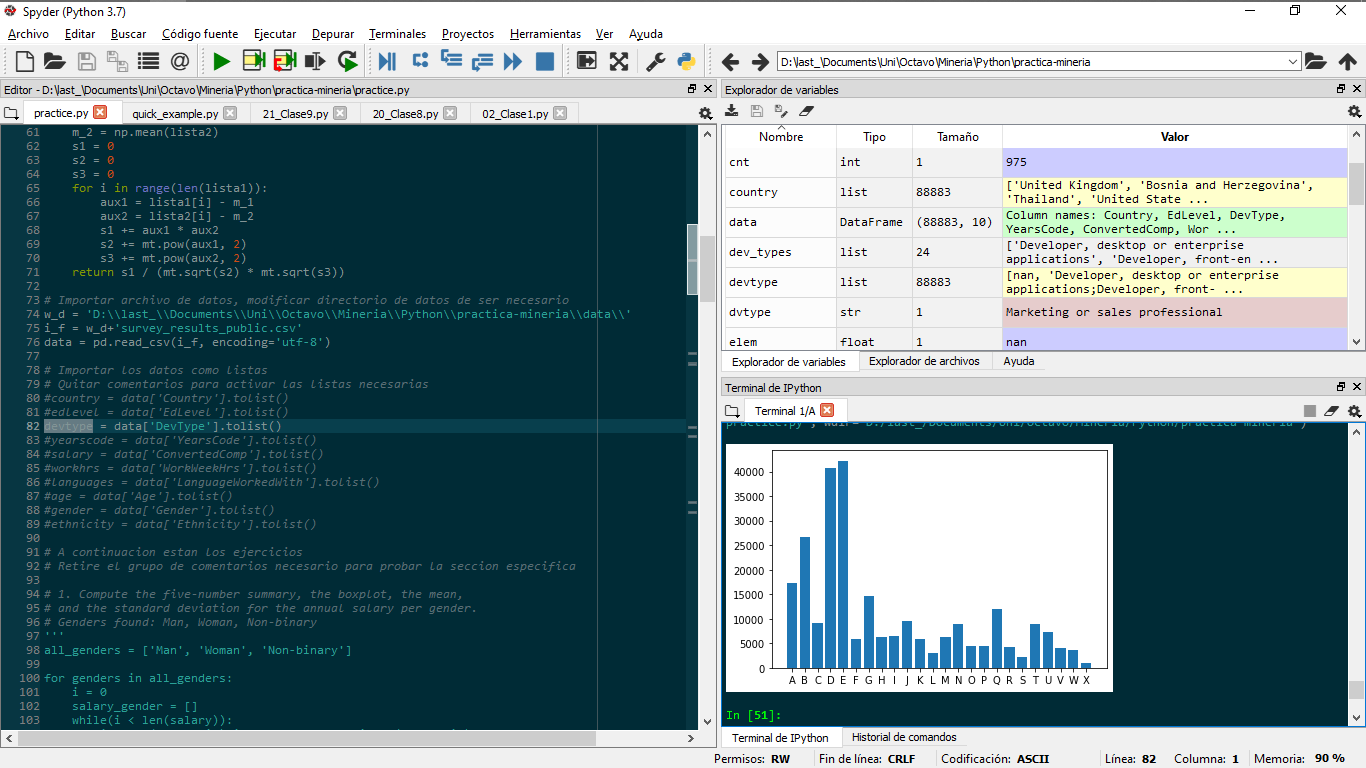
Hay una gran cantidad de países en la lista, así que analizar los datos referentes al salario implica revisar una lista bastante larga. Por cuestiones de espacio, en vez de incluir capturas de pantalla de los resultados, estos se ubican en "./data/output\_salary\_country.txt", relativo a donde se encuentre la aplicación al momento de ejecutarla, por si desea revisarlos.

* El país mejor pagado es, aparentemente, Reino Unido.
* Existen varios países en los que es imposible realizar correctamente el análisis. Esto se debe a que, o no existen datos referentes al salario de dichas personas, o simplemente no existe el suficiente número de personas para analizar correctamente a dichas naciones. Ejemplos de estos casos son Chad, Liberia y Niger, por dar algunos ejemplos.
* El país con menor dispersión es Monaco. Esto es posible que se deba a la carencia de entrevistados provenientes de dicha nación.
* Según los resultados, las naciones peor pagadas son Burundi, Haití y las Maldivas.

1. **Obtener un bar-plot con las frecuencias de respuestas de cada tipo de desarrollador.**

Para este ejercicio, con tal de evitar errores en los nombres de cada barra, se asignaron letras del alfabeto. Por ejemplo, la primera categoría 'Developer, desktop or enterprise applications' corresponde a la ‘A’, 'Developer, front-end' corresponde a la ‘B’ y así sucesivamente. El orden existe de acuerdo con la siguiente lista:

['Developer, desktop or enterprise applications', 'Developer, front-end', 'Designer', 'Developer, back-end', 'Developer, full-stack', 'Academic researcher', 'Developer, mobile', 'Data or business analyst', 'Data scientist or machine learning specialist', 'Database administrator', 'Engineer, data', 'Engineer, site reliability', 'Developer, QA or test', 'DevOps specialist', 'Developer, game or graphics', 'Educator', 'Student', 'Engineering manager', 'Senior executive/VP', 'System administrator', 'Developer, embedded applications or devices', 'Product manager', 'Scientist', 'Marketing or sales professional']

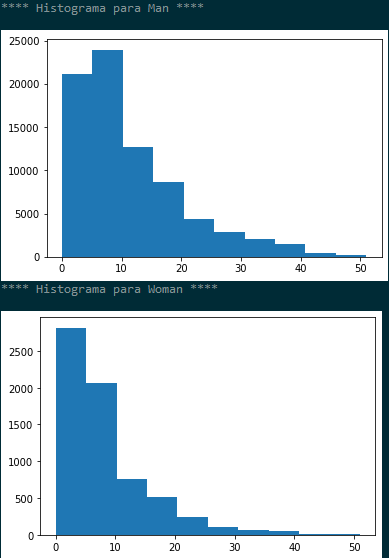
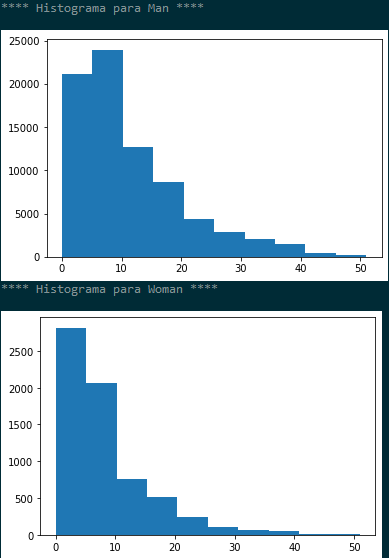


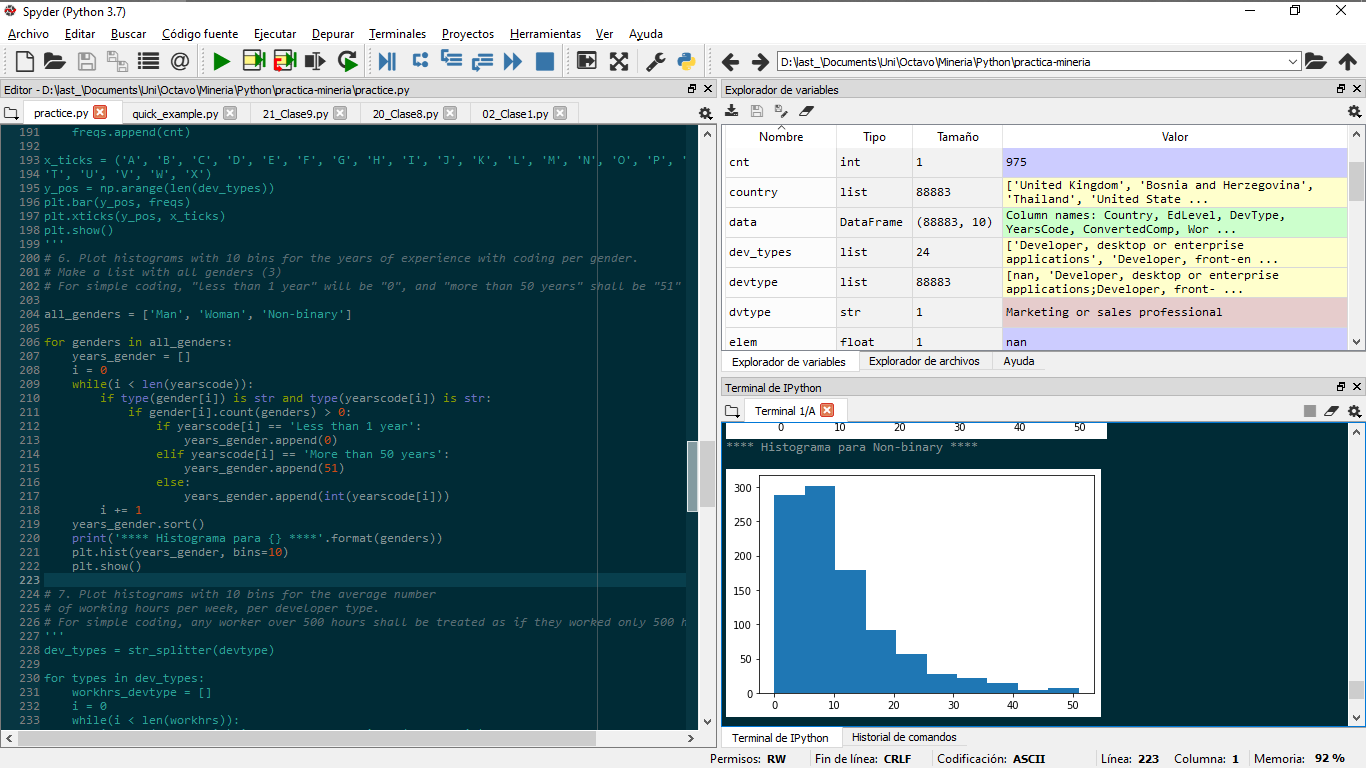
* La mayoría de los entrevistados son programadores full-stack, seguidos por programadores de back-end y de front-end, respectivamente.
* El puesto menos popular se compone de los encargados de ventas o marketing, seguidos por los VP y los ingenieros de confiabilidad.
* Las profesiones más populares, fuera de aquéllas con “programador” en el nombre, son especialistas en DevOps, estudiantes y administradores de bases de datos.

1. **Graficar histogramas con 10 bins para los años de experiencia en programación por género.**

Debido a que solo existen tres opciones de género en la encuesta, revisar los resultados de los histogramas fue relativamente sencillo.

Con tal de simplificar la lectura de resultados y la codificación, a aquéllos con menos de un año de experiencia se les colocó como ‘0’ en los resultados, y como ‘51’ a aquéllos con más de 50 años.

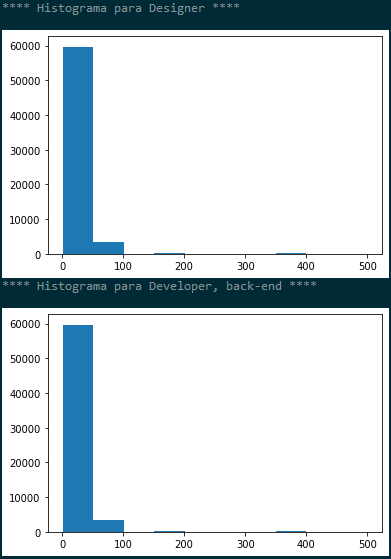
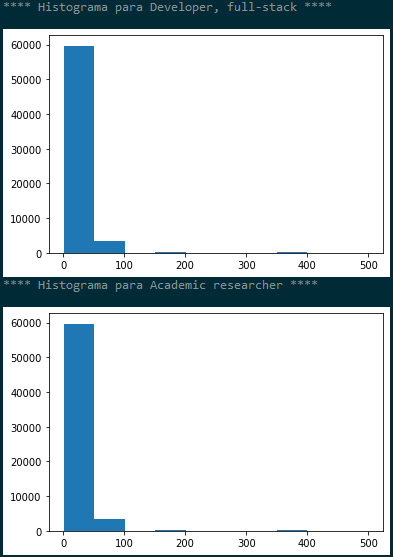
 

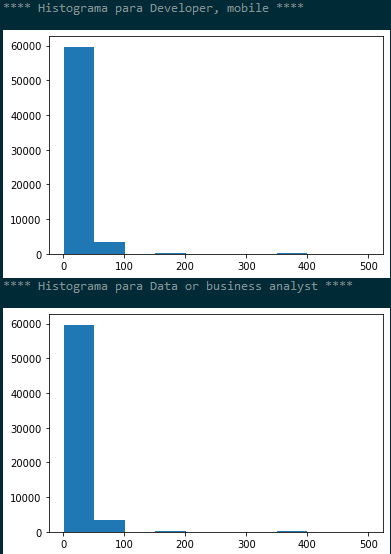
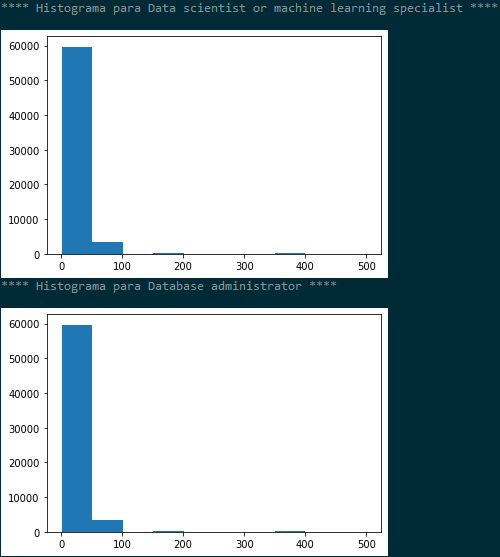


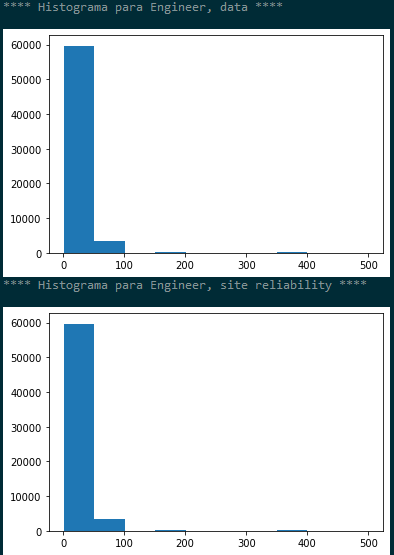
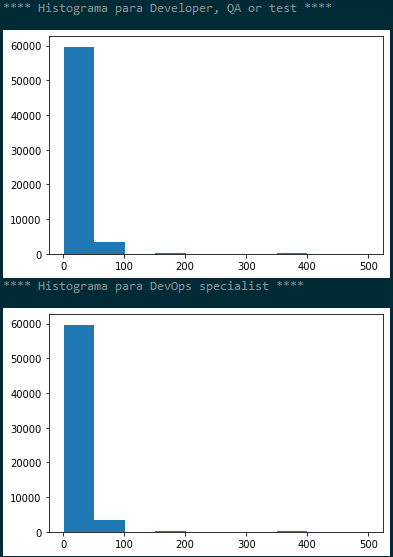
* La mayoría de las personas llevan entre menos de un año de experiencia y 10 años, independiente de su género.
* La mayoría de entrevistados son hombres mientras la minoría son personas no-binarias.
* Desde los 10 años en adelante, la distribución aparenta ser similar para todos los géneros.
* En el caso de las mujeres, hay una ligera depresión en el tercer bin con respecto a los otros dos histogramas, y el primer bin es mayor al segundo. Fuera de esas ligeras diferencias, el punto anterior sigue aplicando.

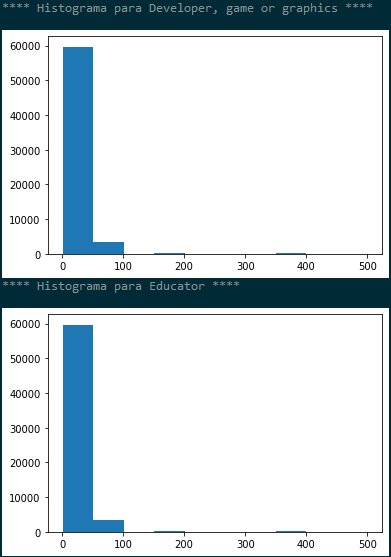
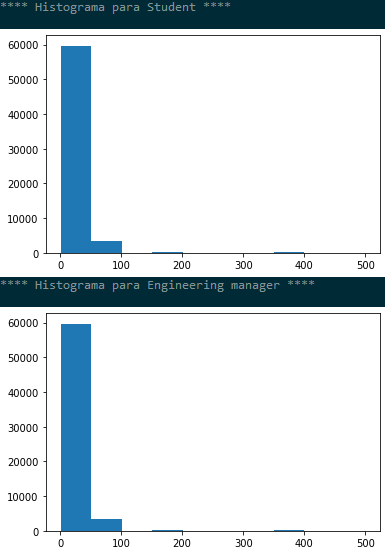
1. **Graficar histogramas con 10 bins para el número de horas de trabajo por semana, por tipo de desarrollador.**

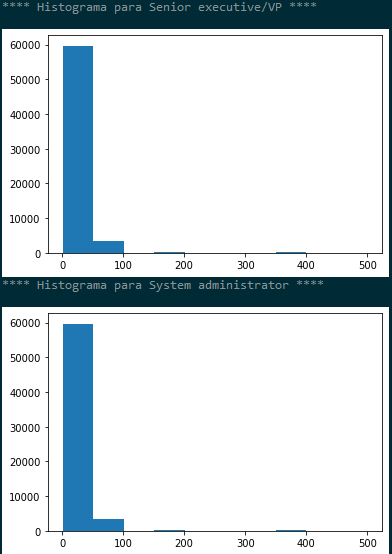
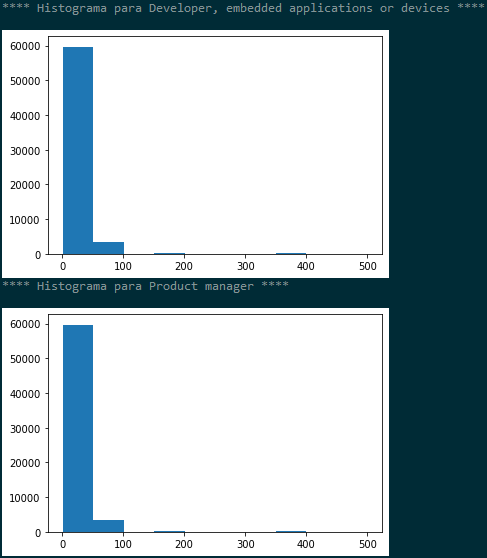
Dentro de los datos existen personas cuyos horarios van hasta ~4000 horas, por lo que se simplifican dichos datos para que el máximo a tomar en cuenta sean 500 horas.

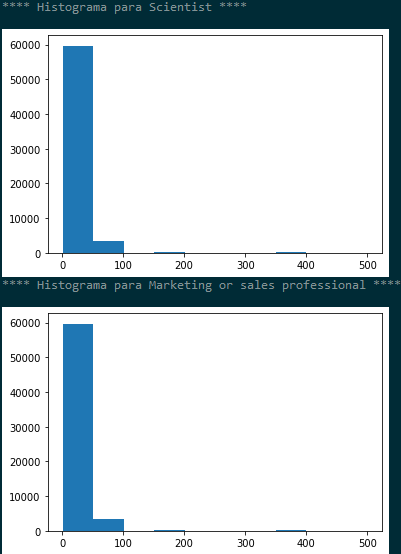
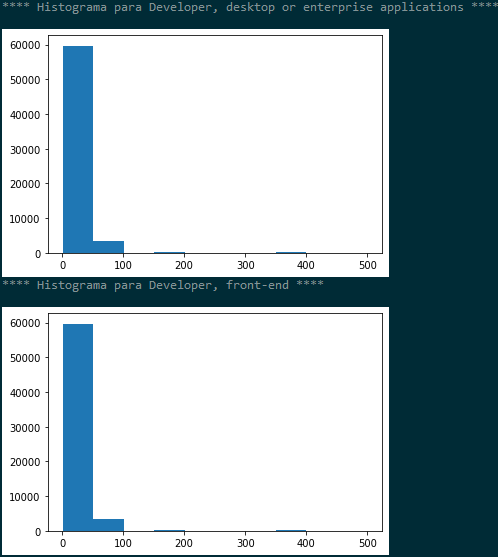
 

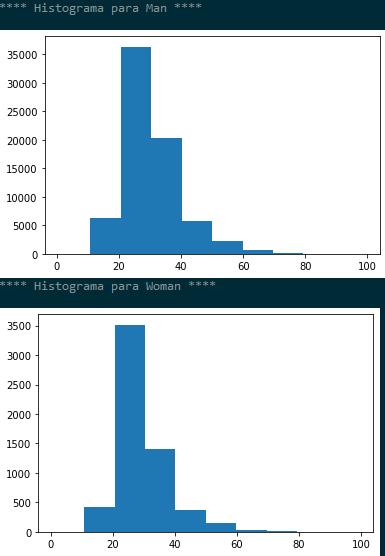
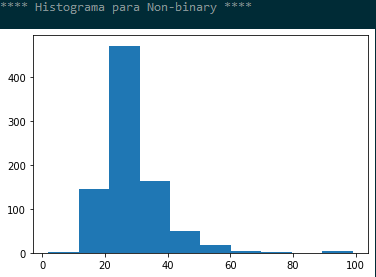
 

* Todos los histogramas son idénticos, por lo que las horas de trabajo para cada tipo de desarrollador son aparentemente las mismas.
* La mayoría de los desarrolladores trabajan 50 horas por semana o menos, con muy pocas personas yendo por encima de dicho tiempo.
* Existe una base de usuarios que trabajan entre 150 y 200 horas, y entre 350 y 400. Aunque sea pequeña, esta sobresale por poco del resto de los histogramas.

1. **Graficar histogramas con 10 bins para la edad por género.**

* De forma similar al ejercicio anterior, el histograma es similar para cada género. La diferencia radica en la cantidad de personas consideradas para cada género.
* La mayoría de entrevistados son personas jóvenes, de entre 20 y 30 años aproximadamente. La mayoría de entrevistados ronda el intervalo de entre 15 y 40 años.
* Aunque poco notable, existe una muy pequeña base de usuarios cuya edad es menor a los 15 años, así como hay personas que se aproximan a los 100 años.
* Los programadores mayores a 40 años son muy raros de encontrar, según los datos.

1. **Calcular la media, mediana y desviación estándar de edades por lenguaje de programación.**

A continuación, se incluye la salida de la aplicación, indicando los resultados del análisis:

\*\*\*\* Resumen de estadistica para HTML/CSS \*\*\*\*

Media: 29.83861149315314

Mediana: 16.0

Desviacion estandar: 8.837072940185127

\*\*\*\* Resumen de estadistica para Java \*\*\*\*

Media: 29.963423514698665

Mediana: 20.0

Desviacion estandar: 8.758582417145659

\*\*\*\* Resumen de estadistica para JavaScript \*\*\*\*

Media: 30.1186744966443

Mediana: 35.5

Desviacion estandar: 8.676488395662691

\*\*\*\* Resumen de estadistica para Python \*\*\*\*

Media: 29.330333648768278

Mediana: 31.0

Desviacion estandar: 9.119177397416664

\*\*\*\* Resumen de estadistica para C++ \*\*\*\*

Media: 28.658021405201577

Mediana: 21.0

Desviacion estandar: 9.837132003254386

\*\*\*\* Resumen de estadistica para C \*\*\*\*

Media: 30.095139607032056

Mediana: 25.0

Desviacion estandar: 9.118472430968199

\*\*\*\* Resumen de estadistica para C# \*\*\*\*

Media: 30.635184057794554

Mediana: 32.0

Desviacion estandar: 9.204646492094136

\*\*\*\* Resumen de estadistica para SQL \*\*\*\*

Media: 30.786009568489014

Mediana: 32.5

Desviacion estandar: 9.002061155682236

\*\*\*\* Resumen de estadistica para VBA \*\*\*\*

Media: 33.40846573393408

Mediana: 28.0

Desviacion estandar: 11.527017473084102

\*\*\*\* Resumen de estadistica para R \*\*\*\*

Media: 30.32078066618968

Mediana: 33.0

Desviacion estandar: 8.780316618709092

\*\*\*\* Resumen de estadistica para Bash/Shell/PowerShell \*\*\*\*

Media: 31.12244541185343

Mediana: 44.0

Desviacion estandar: 9.190332182986298

\*\*\*\* Resumen de estadistica para Ruby \*\*\*\*

Media: 30.89222572414828

Mediana: 25.0

Desviacion estandar: 8.480707415815473

\*\*\*\* Resumen de estadistica para Rust \*\*\*\*

Media: 28.045826645264853

Mediana: 26.0

Desviacion estandar: 8.249230627169382

\*\*\*\* Resumen de estadistica para TypeScript \*\*\*\*

Media: 29.97613629664564

Mediana: 27.0

Desviacion estandar: 7.662254027029062

\*\*\*\* Resumen de estadistica para WebAssembly \*\*\*\*

Media: 30.63202846975089

Mediana: 35.0

Desviacion estandar: 10.042090471351964

\*\*\*\* Resumen de estadistica para Other(s): \*\*\*\*

Media: 32.80376162299239

Mediana: 33.5

Desviacion estandar: 11.114073285321599

\*\*\*\* Resumen de estadistica para Go \*\*\*\*

Media: 30.384147850293843

Mediana: 28.5

Desviacion estandar: 8.293770201413398

\*\*\*\* Resumen de estadistica para PHP \*\*\*\*

Media: 29.400688317983516

Mediana: 35.5

Desviacion estandar: 8.817271681658982

\*\*\*\* Resumen de estadistica para Assembly \*\*\*\*

Media: 28.45338237932891

Mediana: 24.0

Desviacion estandar: 10.574072884286055

\*\*\*\* Resumen de estadistica para Kotlin \*\*\*\*

Media: 28.734044655206482

Mediana: 40.0

Desviacion estandar: 7.864521375367856

\*\*\*\* Resumen de estadistica para Swift \*\*\*\*

Media: 30.182122465052174

Mediana: 28.0

Desviacion estandar: 9.226921719994417

\*\*\*\* Resumen de estadistica para Objective-C \*\*\*\*

Media: 31.783998913338767

Mediana: 38.0

Desviacion estandar: 9.40456017301313

\*\*\*\* Resumen de estadistica para Elixir \*\*\*\*

Media: 31.07210338680927

Mediana: 25.5

Desviacion estandar: 9.132075633966494

\*\*\*\* Resumen de estadistica para Erlang \*\*\*\*

Media: 31.720121028744327

Mediana: 29.0

Desviacion estandar: 9.61715815063997

\*\*\*\* Resumen de estadistica para Clojure \*\*\*\*

Media: 32.46028037383178

Mediana: 20.0

Desviacion estandar: 9.358395775106152

\*\*\*\* Resumen de estadistica para F# \*\*\*\*

Media: 31.884883720930233

Mediana: 22.0

Desviacion estandar: 10.411420449307627

\*\*\*\* Resumen de estadistica para Scala \*\*\*\*

Media: 31.169406239516942

Mediana: 31.0

Desviacion estandar: 8.547448495795111

\*\*\*\* Resumen de estadistica para Dart \*\*\*\*

Media: 27.748502994011975

Mediana: 32.0

Desviacion estandar: 8.78752876934099

* Aparentemente, las personas más jóvenes prefieren usar HTML/CSS.
* Tanto la media como la mediana de edad por cada lenguaje de programación ronda entre los 20 y los 30 años. Esto es coherente con los resultados del ejercicio anterior, pues indica que los programadores son personas jóvenes.
* Existe menor dispersión en los datos referentes a TypeScript, de entre el resto de los lenguajes. Muy de cerca le sigue Kotlin.
* Las personas mayores aparentemente prefieren a Bash, Kotlin o PHP sobre el resto de los lenguajes de programación.

1. **Calcular la correlación entre los años de experiencia y el salario anual.**

De forma similar a como se aplicó el ejercicio 6, a personas con más de 50 años de experiencia se les otorga el valor 51, mientras que aquéllos con menos de un año tendrán valor 0. Esto con tal de facilitar el cálculo de la correlación.

También cabe mencionar que se requirió recortar ambas listas a la par para que incluyeran la misma cantidad de datos, pues de lo contrario habría sido imposible calcular la correlación. Esto se logró eliminando a la par cada dato *n* que fuese vacío de ambas columnas. Esto nos dejó con dos listas con 55737 datos.

Salida del programa:

La correlacion entre el salario y los anios de experiencia es: 0.10598716593336449

Esto indica una ligera proporcionalidad entre el salario anual y los años de experiencia de un programador: a más experimentado, mayor es el salario que éste gana.

1. **Calcular la correlación entre la edad y el salario anual.**

De forma similar al ejercicio anterior, se purgan los datos nulos en ambas tablas hasta tener únicamente 53513 datos.

Salida del programa:

La correlacion entre el salario y la edad es: 0.10826846489974247

* Nuevamente se encuentra una ligera proporcionalidad entre el salario anual y la edad del programador: entre mayor es el individuo, mayor es su salario.
* Desde cierto punto de vista, esta información es congruente con el ejercicio anterior, pues es más probable que una persona más experimentada sea mayor de edad y, por tanto, que tenga un mayor salario, que el resto de los individuos.

1. **Calcular la correlación entre el nivel educativo y el salario anual.**

Para realizar este ejercicio fue necesario asignar índices ordinales a los diferentes niveles con tal de así tener valores para calcular de forma efectiva la correlación. A aquéllos que nunca completaron algún nivel se les otorgó el 0, a la primaria se le otorgó 1, a la secundaria 2 y así sucesivamente.

De forma parecida a los ejercicios anteriores, fue necesario eliminar los datos vacíos de ambas tablas, hasta que ambas tuvieran exactamente 54992 datos.

Salida del programa:

La correlacion entre el salario y el nivel educativo es: 0.03076144020346049

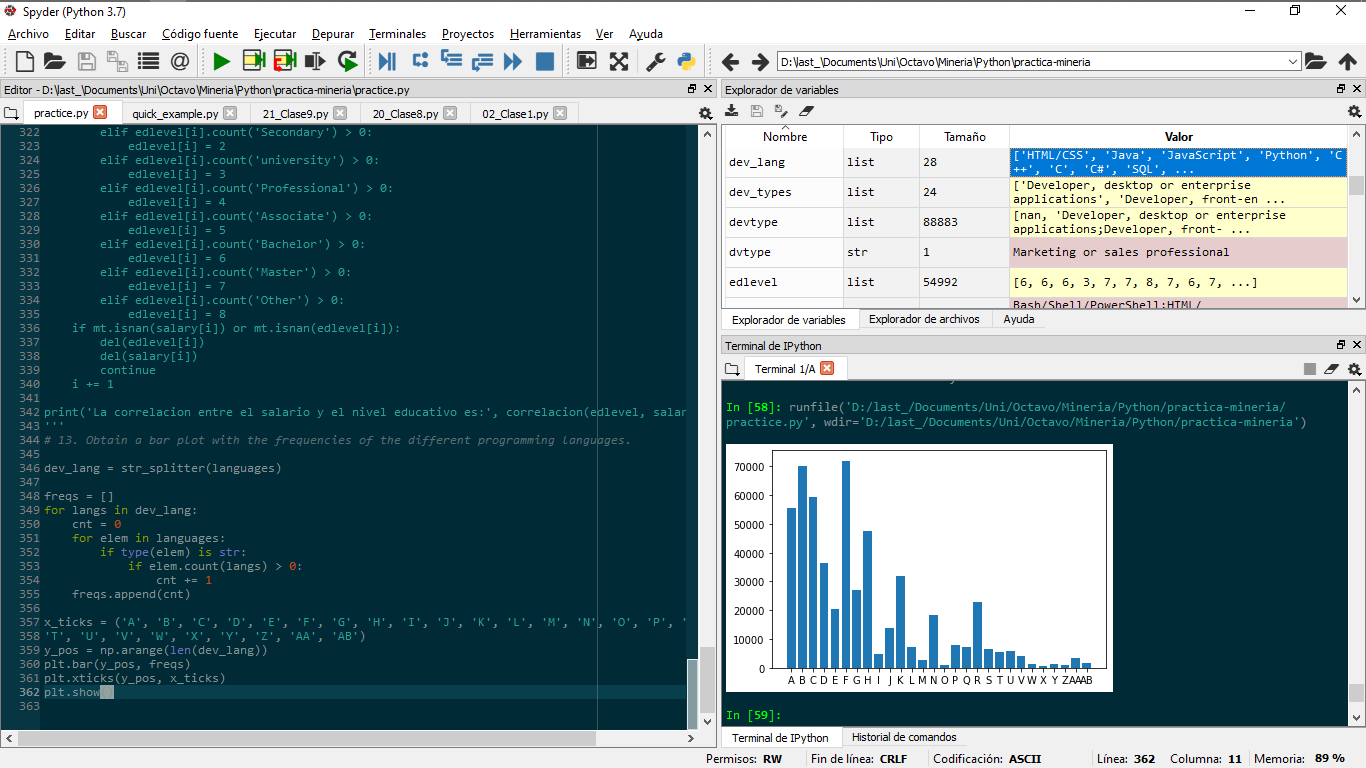
* Nuevamente, existe una muy ligera proporcionalidad entre el nivel educativo y el salario anual de los individuos: entre más lejos llegasen en su educación mayor es el salario que reciben.
* Este resultado es congruente con los ejercicios anteriores, pues es bastante probable que una persona mayor de edad tenga una vasta experiencia, y haya completado su educación al máximo grado posible. Por lo tanto, tendría sentido si sus ganancias anuales fueran mayores al resto de los entrevistados.

1. **Obtener un bar-plot con las frecuencias de los diferentes lenguajes de programación.**

De forma similar al ejercicio 5 y con tal de evitar relleno inesperado en las etiquetas, se usaron las letras del alfabeto para así evitar la sobreposición entre etiquetas, y así poder analizar la gráfica sin inconvenientes.

Cada letra indica, en orden, los siguientes lenguajes de programación:

['HTML/CSS', 'Java', 'JavaScript', 'Python', 'C++', 'C', 'C#', 'SQL', 'VBA', 'R', 'Bash/Shell/PowerShell', 'Ruby', 'Rust', 'TypeScript', 'WebAssembly', 'Other(s):', 'Go', 'PHP', 'Assembly', 'Kotlin', 'Swift', 'Objective-C', 'Elixir', 'Erlang', 'Clojure', 'F#', 'Scala', 'Dart']



* La mayoría de las personas practican ya sea C o Java.
* El lenguaje menos practicado es, aparentemente, Erlang.
* Aunque no como al grado de C o Java, la programación Web es bastante popular, pues una parte importante de los entrevistados practica HTML/CSS y/o JavaScript.
* Aproximadamente una tercera parte (en cantidad) de personas que conocen C también conocen C++ y/o C#.
* SQL es al parecer bastante conocido, pues ocupa el quinto lugar entre los lenguajes más utilizados.