Análisis y Visualización de Datos

Descifrando el Universo: Apariencia de las galaxias

Mentora: Daza Vanessa

Autores: Montoya Gabriel,
Paludi Franco,
Zacco Franco.

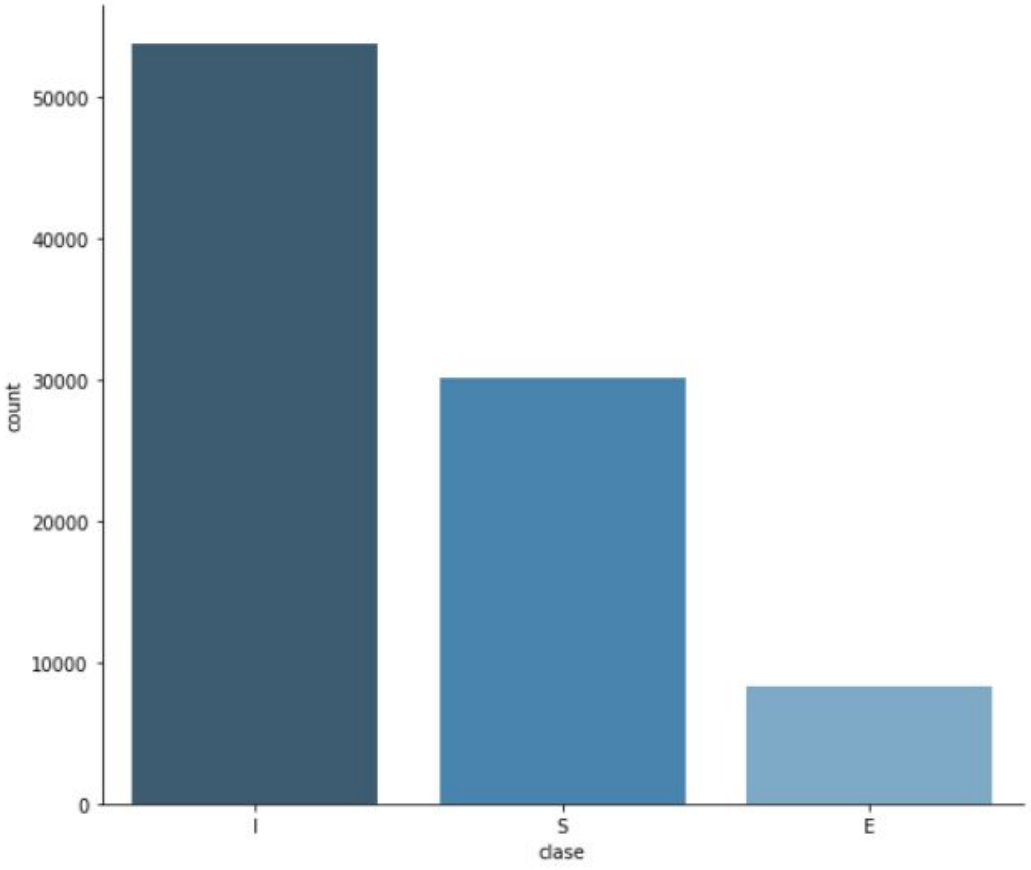
Análisis y Visualización de Datos

Córdoba, 2019

Visualización

- ***Encuentre una forma adecuada para exhibir el balance de las clases elíptica (E), espiral (S) e irregular (I) usando la nueva columna clase.***

- Decidimos que la forma mas representativa para exhibir un balance comparativo es un grafico de barras, el cual muestra el siguiente resultado

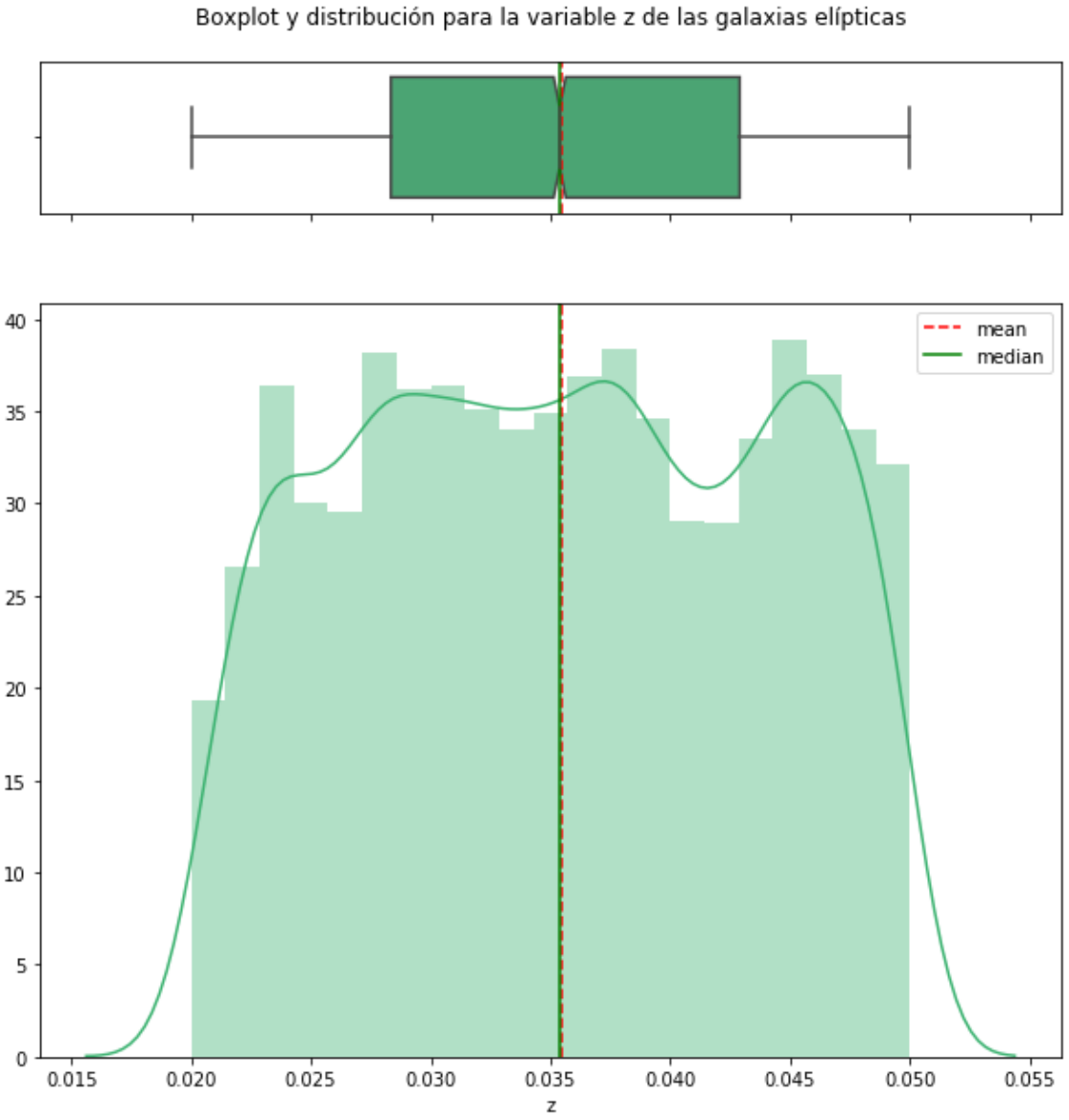


Análisis y Visualización de Datos

- ***Use histogramas y boxplots para visualizar intervalos de confianza, mediana, media, intercuartiles y* outliers *de la posición z y del tamaño R para algún tipo morfológico.***

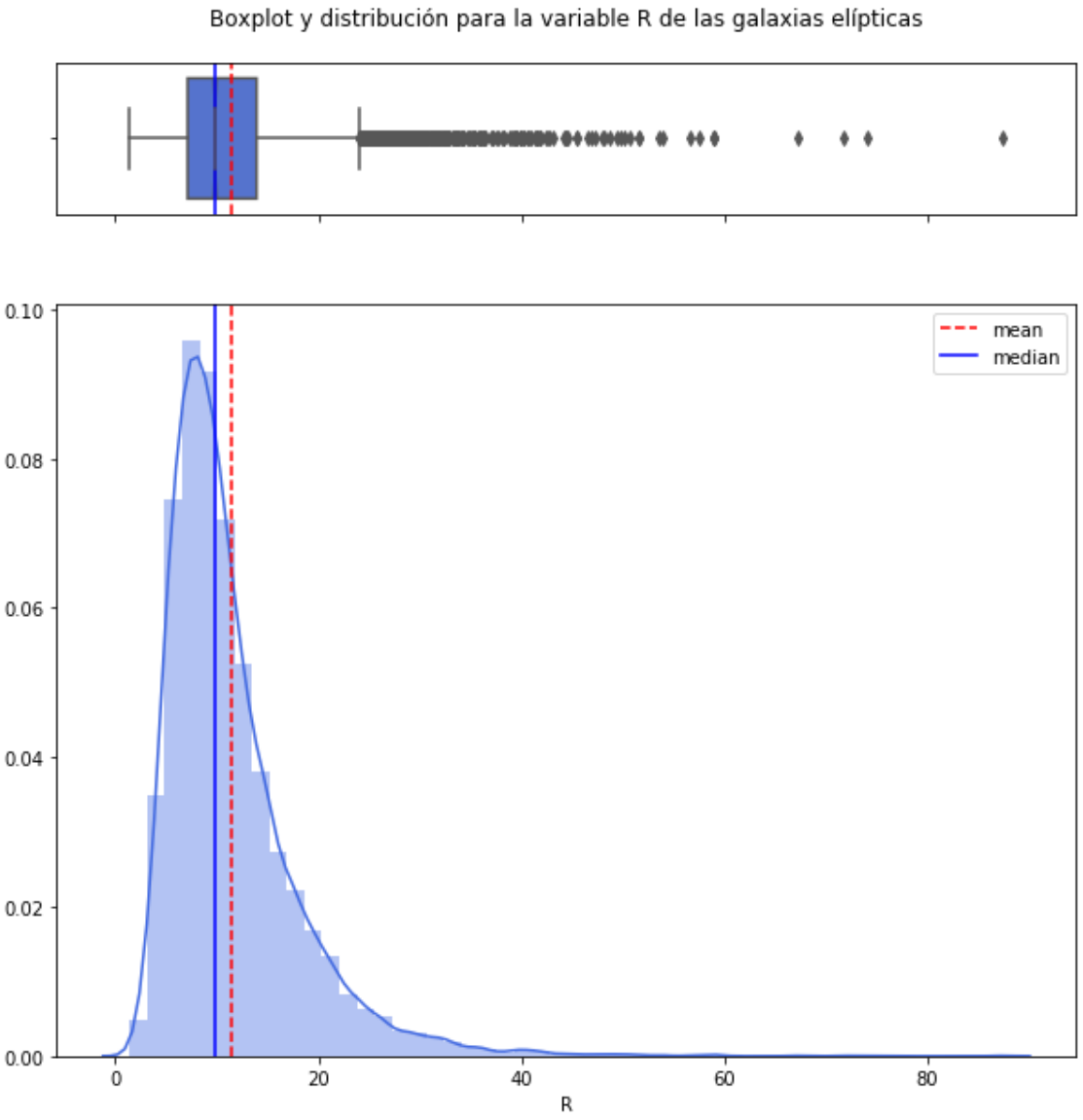
- En este apartado seleccionamos las galaxias elípticas. Dando como resultado los siguientes gráficos:

Boxplot y distribución para la variable z



Análisis y Visualización de Datos

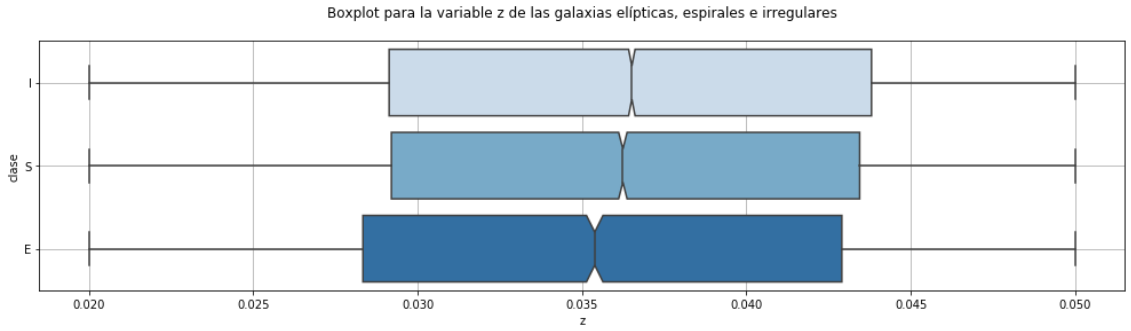
Boxplot y distribución para la variable R



- ***Mostrar si los valores de las medianas de las distribuciones de z para cada clase (E, S e I) son estadísticamente diferentes.***

- Para ello utilizamos un boxplot. Podemos decir con una confianza del 95% que las medianas son diferentes para las galaxias Elípticas respecto de las galaxias Espirales e Irregulares. No podemos decir esto para los otros tipos de galaxias debido a que sus intervalos de confianza se solapan.

Análisis y Visualización de Datos



Valores Característicos

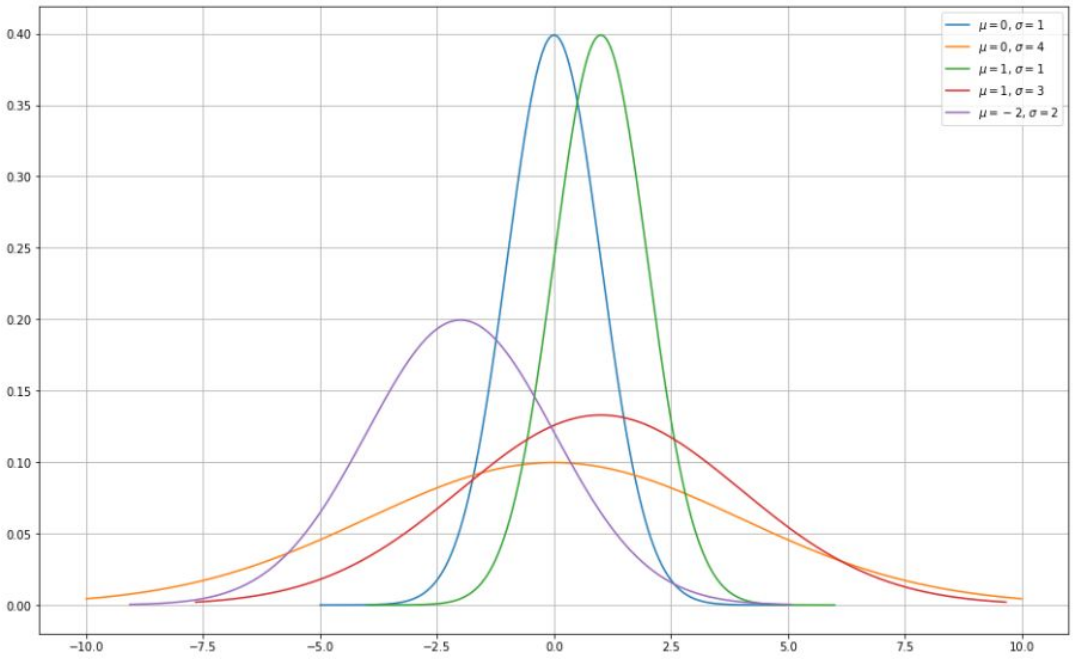
- ***Mostrar los valores explícitos característicos de la distribución z de las galaxias elípticas, espirales, irregulares.***

	Elípticas	**Espirales**	**Irregulares**
count	8257	30046	53799
mean	0.035441	0.036080	0.036198
std	0.008377	0.008375	0.008473
min	0.020001	0.020001	0.020002
25%	0.028338	0.029204	0.029136
50%	0.035387	0.036243	0.036516
75%	0.042901	0.043443	0.043803
max	0.049999	0.049999	0.050000
median	0.035387	0.036243	0.036516

Distribución Normal

- ***Grafique la distribución Normal variando sus estadísticos. Realice una breve descripción de los cambios que nota en estos.***

Análisis y Visualización de Datos



- En el gráfico anterior se observan diferentes distribuciones normales con distintas medias (μ) y desviaciones estandares (σ). Se puede ver que la variación de la media desplaza la campana sobre el eje horizontal, mientras que la desviación estándar nos indica que tan ancha o distribuida va a ser la campana, cuanto más grande el σ más ancha es la campana

Valores Faltantes

- ***Algunos valores físicos de las galaxias no pueden ser determinados. Muchas veces la alternativa al usual* NaN *es llenar la celda del valor faltante con cero o con números de valor muy grande o muy chico discordantes a los valores usuales que toma la cantidad física.***
  Se puede optar por ignorar dicho número o remplazarlo por cero, el valor medio o eliminarlo en el caso que el dataset sea de una gran dimensión.
  El caso anteriormente se observa con la variable* color *donde dos de las galaxias no tienen su color determinado.
  - **Dele solución**
  - **Ajuste la distribución de la variable** color **con el tamaño del bin=0.1 a una Normal usando *sns.distplot***
  - **Explique que observa**

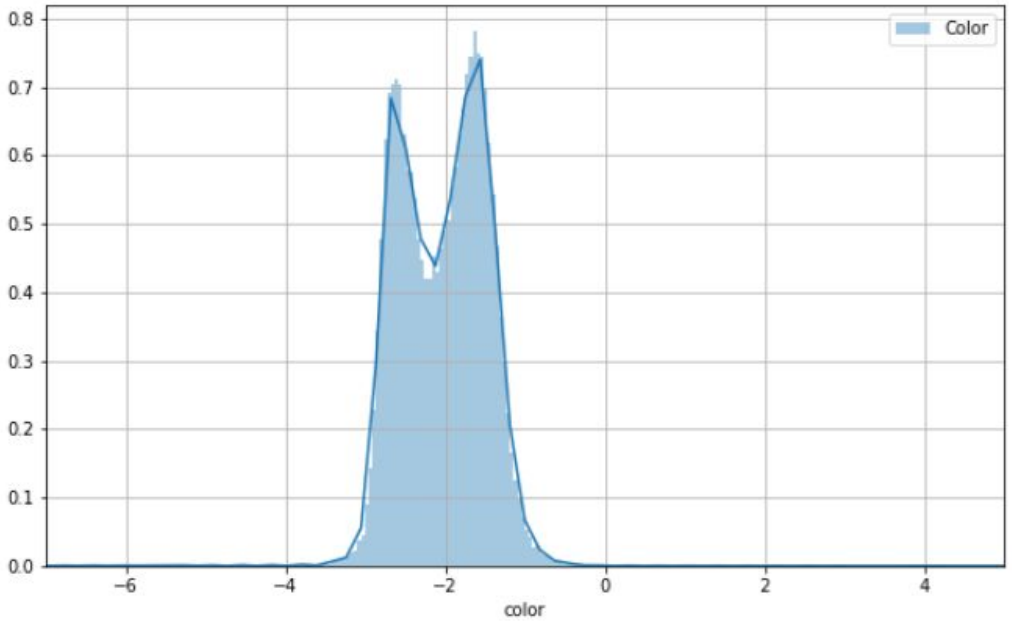
Count	92102
Mean	-1.830732
Std	46.686363
Min	-13.48457
25%	-2.510555
50%	-1.994779

Análisis y Visualización de Datos

75%	-1.606971
Max	10015.86

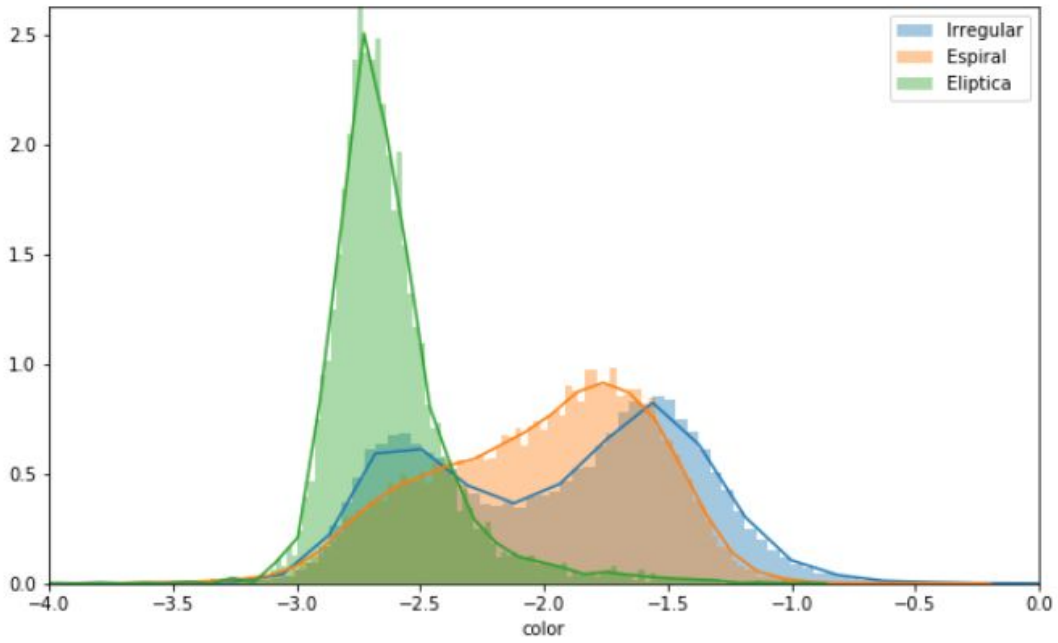
- Se puede ver que el valor máximo dista mucho de la media y los cuartiles. Nos da una pista que puede ser que un valor faltante rellenada con un valor muy grande
- Al buscar las filas con el valor de la variable color mayor a 10 (valor lo suficientemente alejado de la media) el resultado nos devuelve 2 filas. Ambas filas tienen un valor muy grande, las eliminamos ya que el dataset es muy grande y dos resultados menos no impactaran en el análisis.

- ***Ajuste la distribución de la variable color con el tamaño del bin=0.1 a una Normal usando sns.distplot***



- La distribución del color para todas las galaxias parece ser 2 o más distribuciones superpuestas. Vemos como es la distribución del color para las diferentes morfologías.

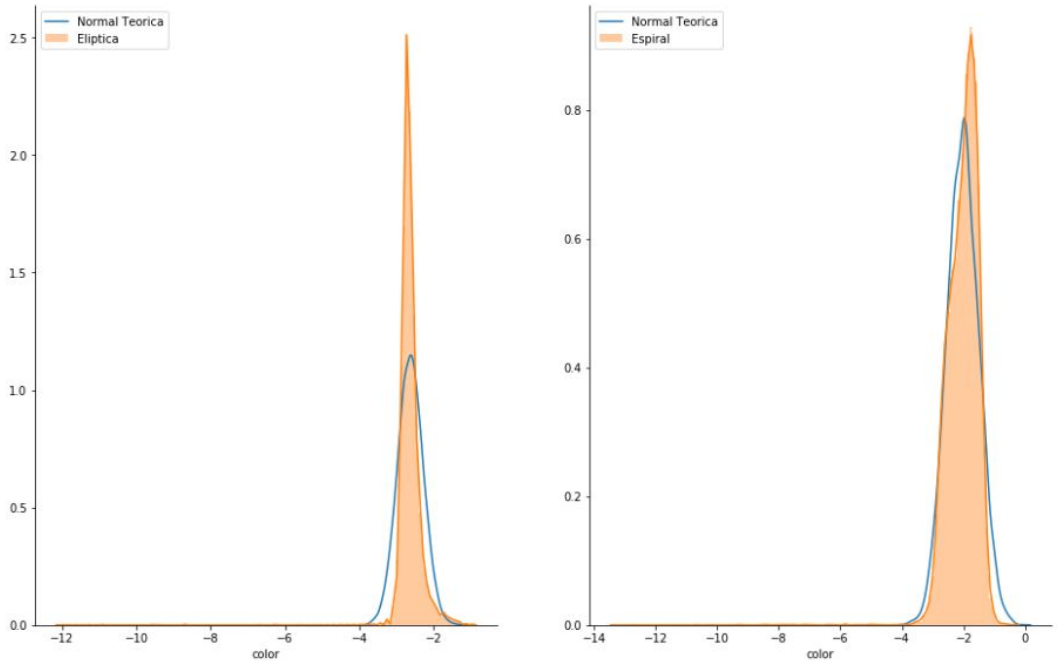
Análisis y Visualización de Datos



- En primer lugar, sabemos que cuanto más negativa es la variable color, más roja es la luz que llega de la galaxia. Por otro lado, conocemos que las estrellas más viejas son más rojas, mientras que las estrellas más jóvenes son más azules.
  Teniendo en cuenta estos 2 puntos y observando el gráfico anterior, podemos decir que en general las galaxias elípticas se formaron en etapas más tempranas del universo, mientras que las espirales lo hicieron más tarde. Otra forma de explicar esta situación es mirando el polvo y gas libre en la galaxia que permite el nacimiento de nuevas estrellas. En las galaxias elípticas hay poco gas y polvo libre, ya que fue utilizado es épocas anteriores para formar estrellas y con esa poca concentración de gas y polvo la tasa de nacimiento de nuevas estrellas es muy baja. En el lado opuesto están las galaxias espirales, donde hay más cantidad de polvo y gas libre que permite el nacimiento de nuevas estrellas.
  Por otro lado en las galaxias irregulares podemos ver tanto galaxias jóvenes como viejas. Este tipo de galaxias suele presentar contenido de gas y polvo suficiente para que se formen nuevas estrellas, aunque algunas no lo hacen siempre, como por ejemplo las galaxias irregulares enanas.

Análisis y Visualización de Datos

- Del gráfico anterior es obvio que la distribución del color no es normal para las galaxias irregulares. Veamos que pasa para las elípticas y las espirales



- A simple vista podemos decir que la distribución del color de las galaxias elípticas no se ajusta muy bien a la normal, pero en las espirales parece haber un mejor ajuste. Los comprobamos utilizando un test de Kolmogorov-Smirnov comparando con la distribución normal. El test nos plantea las siguientes hipótesis:

  H_0: La distribución es Normal

  H_a: La distribución no es Normal

- KS Test Elípticas: p-valor = 6.883279270101887e-180

- KS Test Espirales: p-valor = 4.4647706287249846e-97

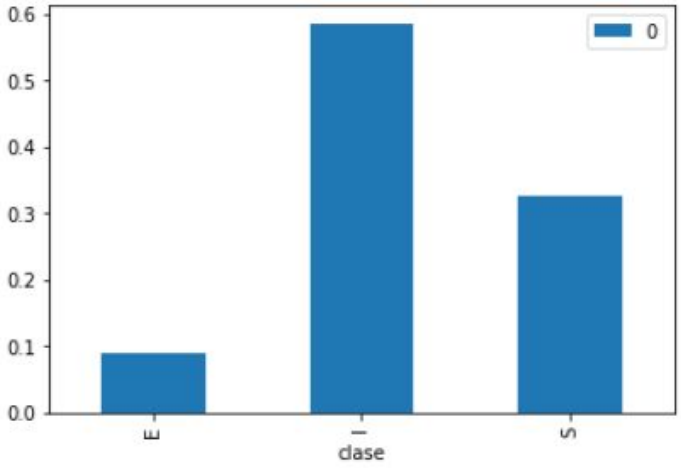
- Viendo que el P-Valor es mucho menor que 0.05 en ambos casos, descartamos la hipótesis nula y decimos ninguna de las distribuciones son normales.

Análisis y Visualización de Datos

Probabilidad

- ***Calcule la Probabilidad marginal de cada tipo morfológico y grafíquela.***

Clase	PM
Elíptica	0.089653
Irregular	0.584115
Espiral	0.326232



- **Calcule la probabilidad conjunta las galaxias de clase = E con el color mayor a -2.1**

  Para calcular esta probabilidad debemos usar el Teorema de Bayes, el cúal nos relaciona las probabilidades conjuntas, condicionales y marginales de la siguiente forma:

  $$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

  Donde:
  - ➔ $P(A/B)$ es la probabilidad condicional
  - ➔ $P(A \cap B)$ es la probabilidad conjunta
  - ➔ $P(B)$ es la probabilidad marginal

  Podemos despejar la probabilidad conjunta de forma que:

  $P(A \cap B)$ = P(A/B) P(B) = P(B/A) P(A)

  Para nuestro caso establecemos que

  - ➔ A el color es mayor a -2,1
  - ➔ B la galaxia es elíptica

Análisis y Visualización de Datos

Del punto anterior, sabemos que P(B) = 0,089653

La probabilidad condicional la obtenemos del dataset y es:

P(A/B) = 0,040451

Así la probabilidad conjunta es:

$$P(A \cap B) = 0,040451 \,.\, 0,089653 = 0.0036264$$

- ***Analice la correlación entre dos de las variables a través de chi-cuadrado. Muestre esta correlación a través de mapas de calor.***

- La correlación utilizando Chi Cuadrado aplica a variables categóricas. Como solo tenemos 1 variable categórica (la clase/morfología de la galaxia) creamos otro que es el nombre del color, que puede tomar valores "rojo" o "azul" de acuerdo al valor de la variable color

Color_name	**Blue**	**Red**
Clase		
Elíptica	656	7601
Irregular	35967	17830
Espiral	21860	8186

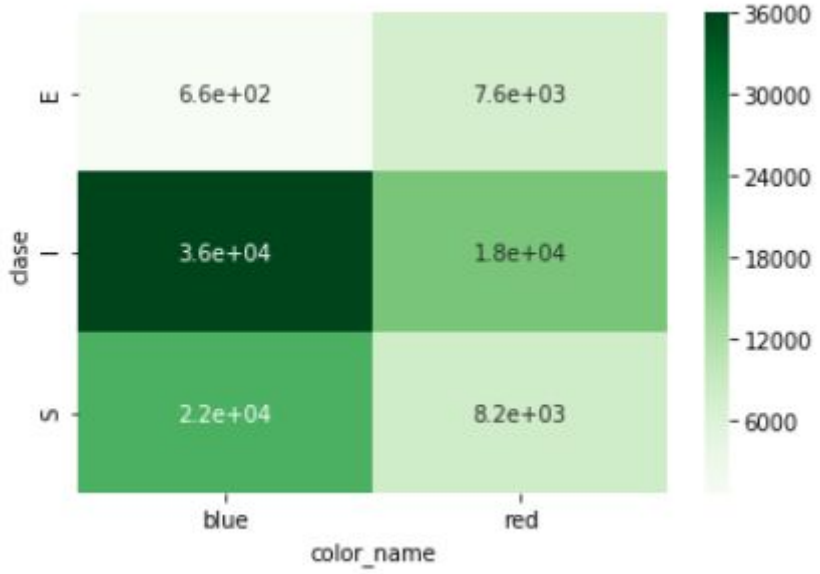
- Calculamos el chi cuadrado para ver la independencia de las variables. Este test plantea las siguientes 2 hipótesis:

  H_0: Las variables son independientes

  H_a: Las variables son dependientes

- El cálculo nos devuelve un p-valor igual a 0. Como el P-Valor es 0, podemos rechazar la hipótesis nula y decimos que las variables son dependientes. Esto es algo esperado, ya que como vimos anteriormente existe una dependencia entre la morfología de las galaxias y el color predominante
- Al graficas las relaciones entre las variables mediante un mapa de calor encontramos el siguiente resultado

Análisis y Visualización de Datos

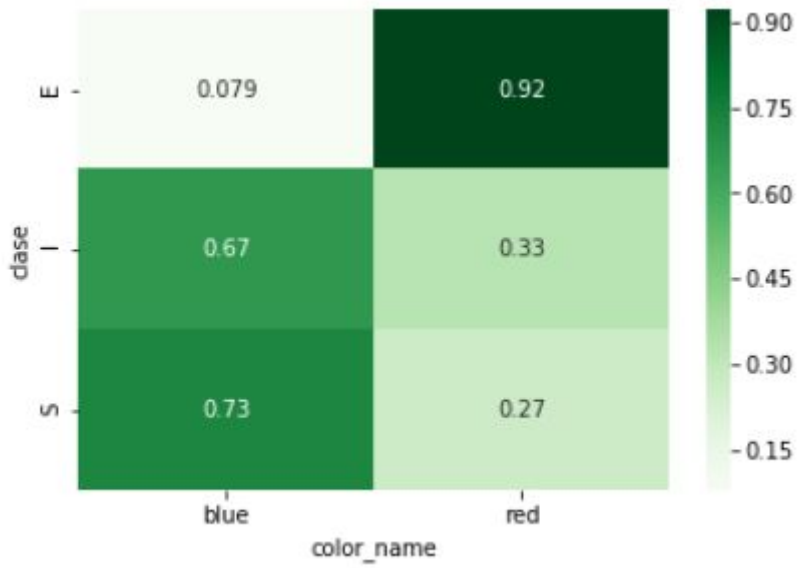


- Un buen recurso para mostrar los resultados de correlaciones suele ser el mapa de calor, pero en este caso resulta engañoso ya que a primera vista nos muestra que existe una correlación débil entre las galaxias elípticas y espirales con los colores de las mismas. Esto se debe a que la probabilidad marginal de ese tipo de galaxias es bajo respecto a las irregulares, que es donde parece haber mayor correlación. Para corregir esto balanceamos las muestras por galaxia, lo cual nos presenta la siguiente tabla:

Color_name	**Blue**	**Red**
Clase		
Elíptica	0.079448	0.920552
Irregular	0.668569	0.331431
Espiral	0.727551	0.272449

- La nueva tabla nos devuelve un nuevo mapa de calor:

Análisis y Visualización de Datos



- De esta forma la situación mejora notoriamente, y el gráfico muestra el resultado correctamente