ה- DataSet מכיל מידע על הכנסות של כ-50,00 אנשים, המידע מכיל נתונים כגון : גיל, סוג משרה, סוג חברה, שנות לימוד, תארים אם קיימים, שעות עבודה בשבוע, צבע העור ועוד...

התיוג שלנו היא העמודה האחרונה, הכנסה שנתית שהיא מתוייגת ע"י "מעל 50K" או "מתחת ל50K-"

**ננסה למצוא תשובה לשאלה – האם צבע העור הינו גורם משפיע על ההכנסה השנתית של האדם.**

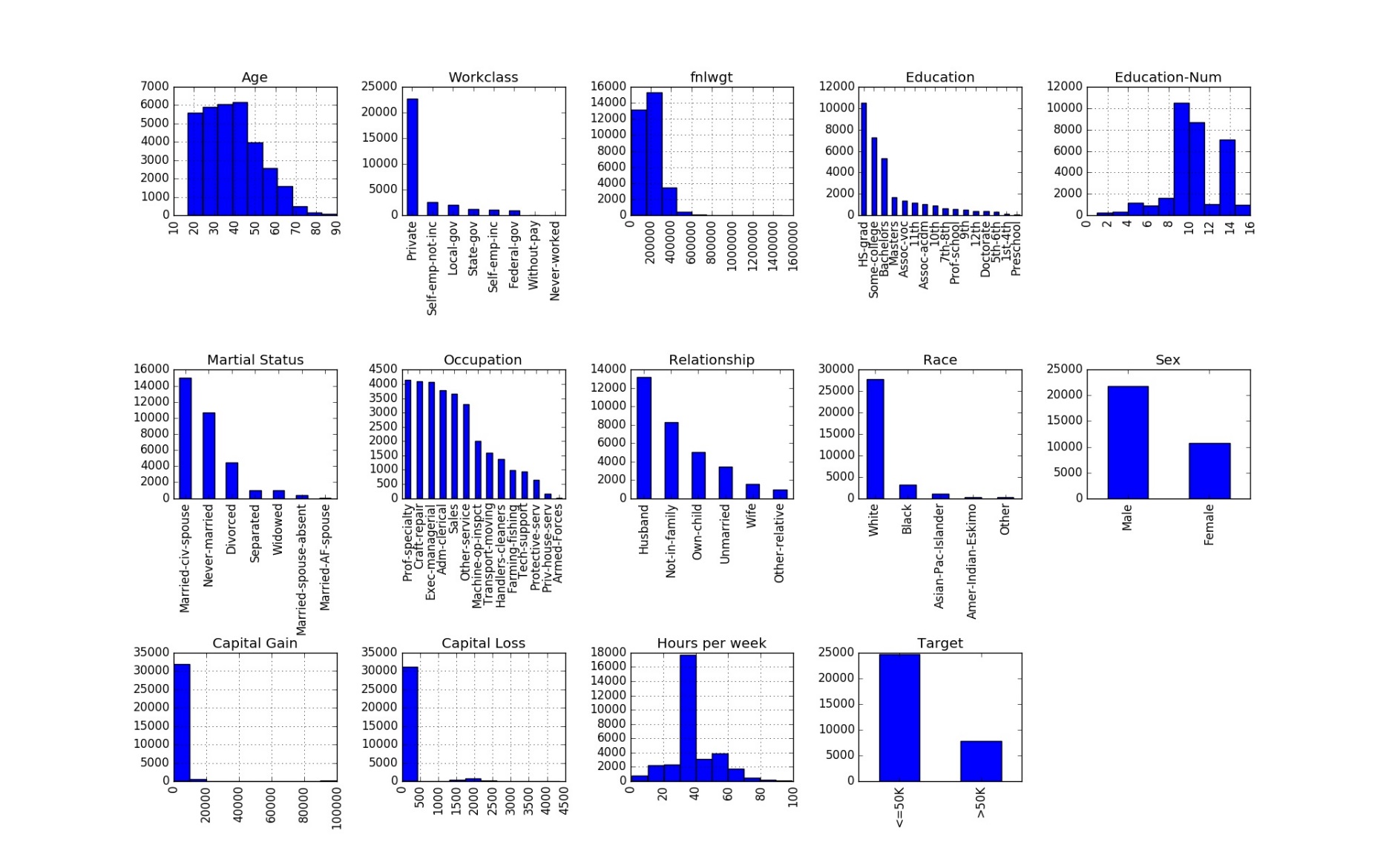
עמודת המדינה הורדה מפני שאינה מעניינת אותנו.

העמודות כרגע הן:

"Age", "Workclass", "fnlwgt", "Education", "Education-Num", "Martial Status",  
"Occupation", "Relationship", "Race", "Sex", "Capital Gain", "Capital Loss", "Hours per week", "Target"

נבחן את המאפיינים ע"י הצגתם בגרף נפרד לכל עמודה

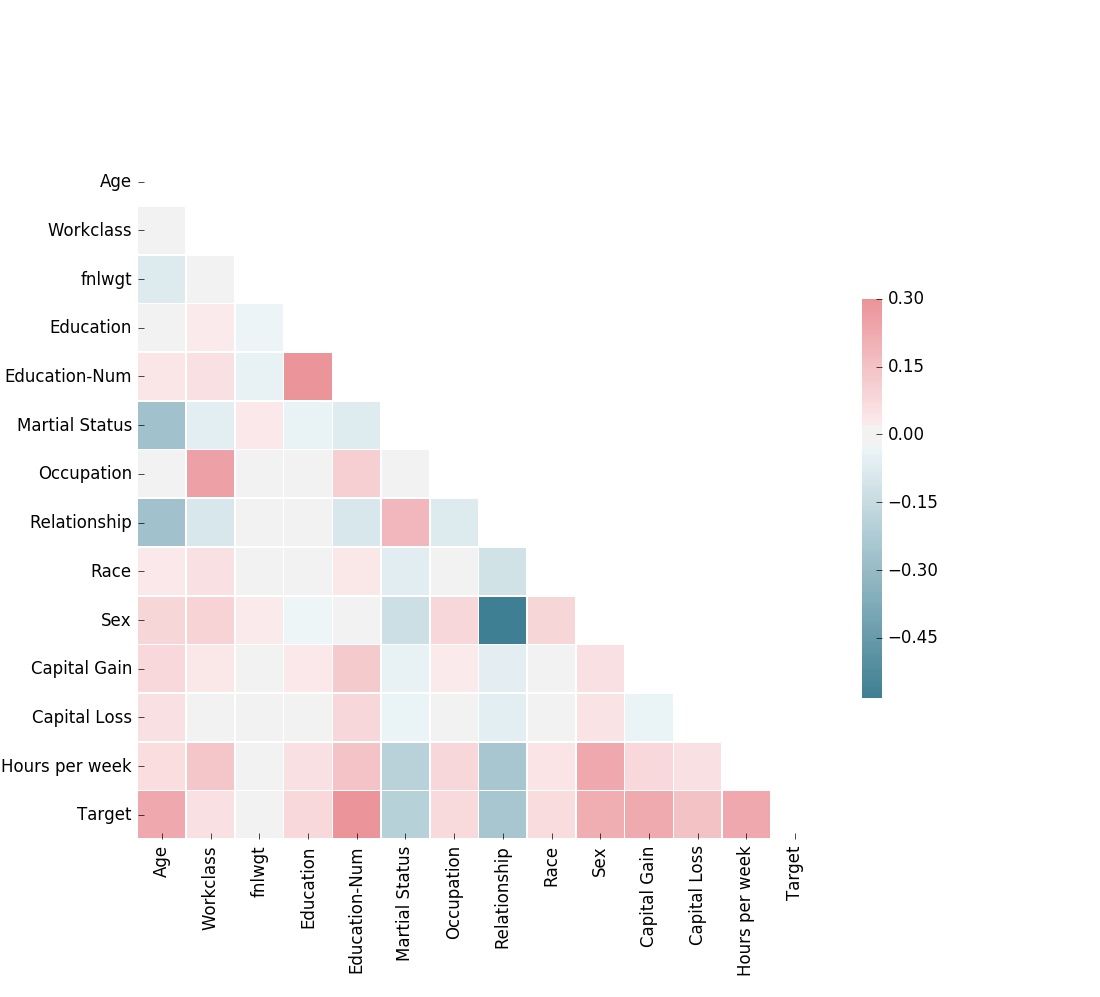
( פונקצייה בשם **plot\_Distributor\_Of\_Each\_Numerical\_Column** ):

ונקבל :

ניתן לראות מהגרפים לעיל שרוב המידע מסתמך על גברים **לבנים** עם הכנסה מתחת ל-50K.

בהמשך, נרצה לנסות ולמצוא את ההקשר שבין המאפיינים השונים, לצורך כך, נצטרך להמיר את הנתונים הקטגוריים לנתונים מספריים (פונקצייה **encode\_Categorical\_Features\_To\_Numerical** ),

לאחר מכן מציג את הקשר בין המאפיינים ע"י מפה ( פונקצייה **plot\_Heat\_Map** ) ונקבל:



הפונקציה מראה רק את המשולש התחתון שכן הקשר בין כל מאפיין לעצמו הוא 1.0 (הגיוני).

ניתן לראות שהקשר הוא החזק ביותר (האדום ביותר) בין שנות הלימוד לבין ההשכלה, מהסתכלות על הנתונים, אנו רואים שעם עלייה במס' שנות הלימוד, כך גם ההשכלה גדלה :



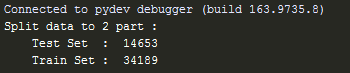
במקרה זה, אין טעם לשני המאפיינים והעדפנו למחוק את עמודת ה"השכלה" ולהשאיר את עמודת שנות הלימוד מפני שהיא מהווה אינדיקציה טובה יותר לרגרסיה שנפעיל בהמשך.

כעת העמודות שלפיהן אנו עובדים הן:

"Age", "Workclass", "fnlwgt", "Education-Num", "Martial Status", "Occupation", "Relationship", "Race", "Sex", "Capital Gain", "Capital Loss", "Hours per week", "Target"

פיצלנו את ה- DATA לשני חלקים ביחס של בערך 1/3 – 2/3 כאשר החלק הגדול יותר מיועד לTraining Set

(פונקציה **split\_Data\_To\_Test\_And\_Train\_Sets** ) ונקבל :



נבחן שוב המאפיינים ע"י הצגתם בגרף נפרד לכל עמודה אך כעת אל מול המידע **המספרי**

( פונקצייה בשם **plot\_Distributor\_Of\_Each\_Numerical\_Column** ):

ונקבל:

