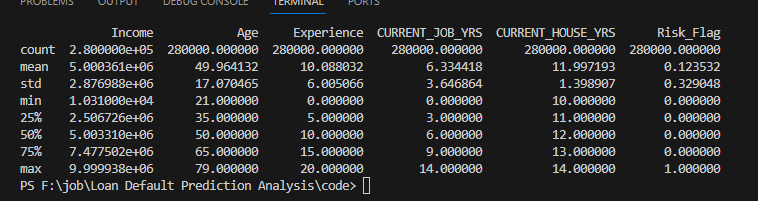
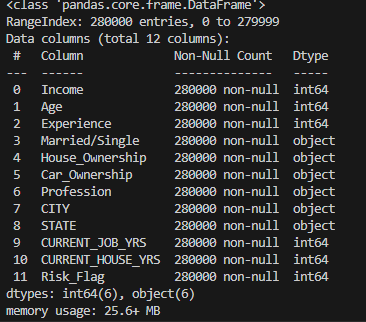
**Data cleaning**

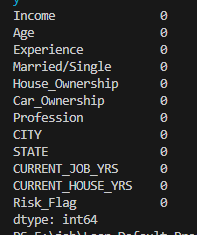
\*کدهای مربوط به این قسمت در فایل data cleaning ذخیره شده است.

1. ابتدا داده های فایل trainو test را ترکیب می کنیم و با نام combined data ذخیره می کنیم.
2. سپس اطلاعات آماری از داده ها را به دست آوردیم.





1. مقادیر missing values را چک می کنیم. مطابق تصویر بالا و پایین هیچ گونه داده nullوجود ندارد.

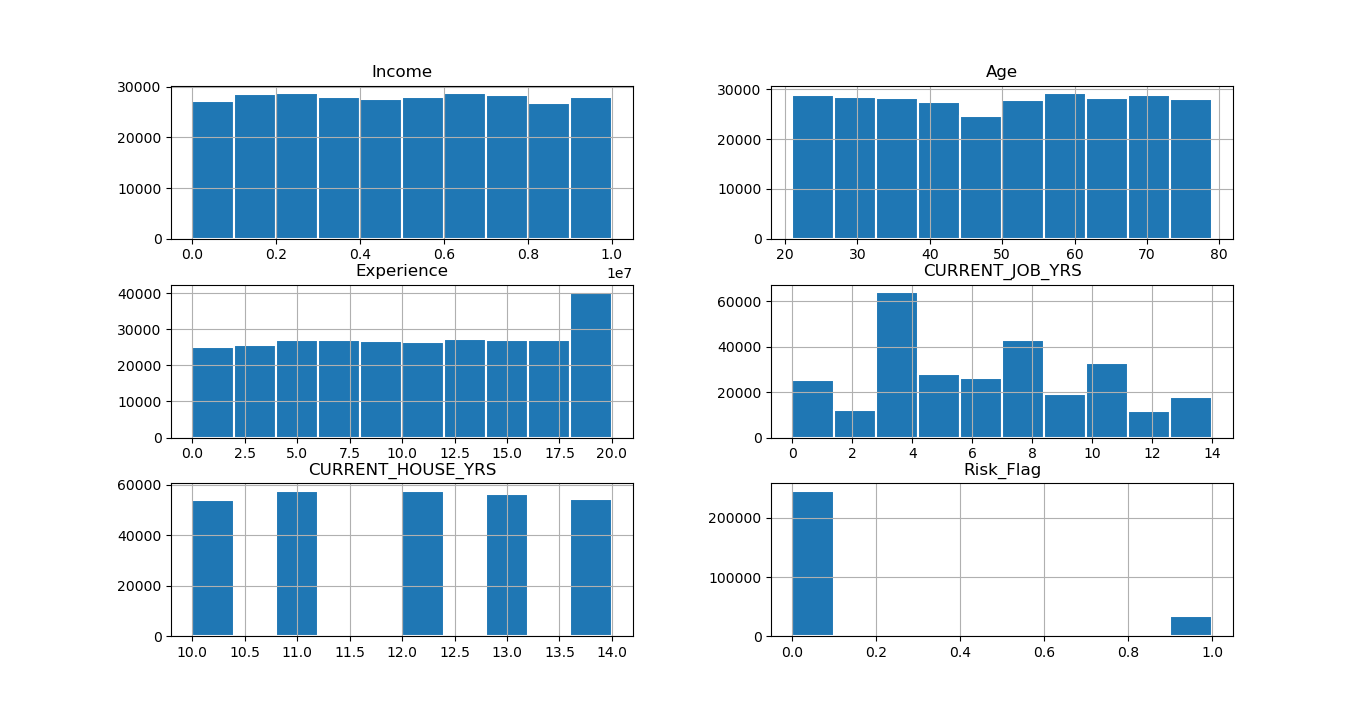


1. در این مرحله موارد تکراری را چک می کنیم. در فایل های training و testنمونه های تکراری وجود ندارند.

**نتیجه گیری**: با توجه به نتایج بالا داده ها تا حد خوبی تمییز هستند و مقادیر از دست رفته یا تکراری و یا نویز وجود ندارد یا اندک است.

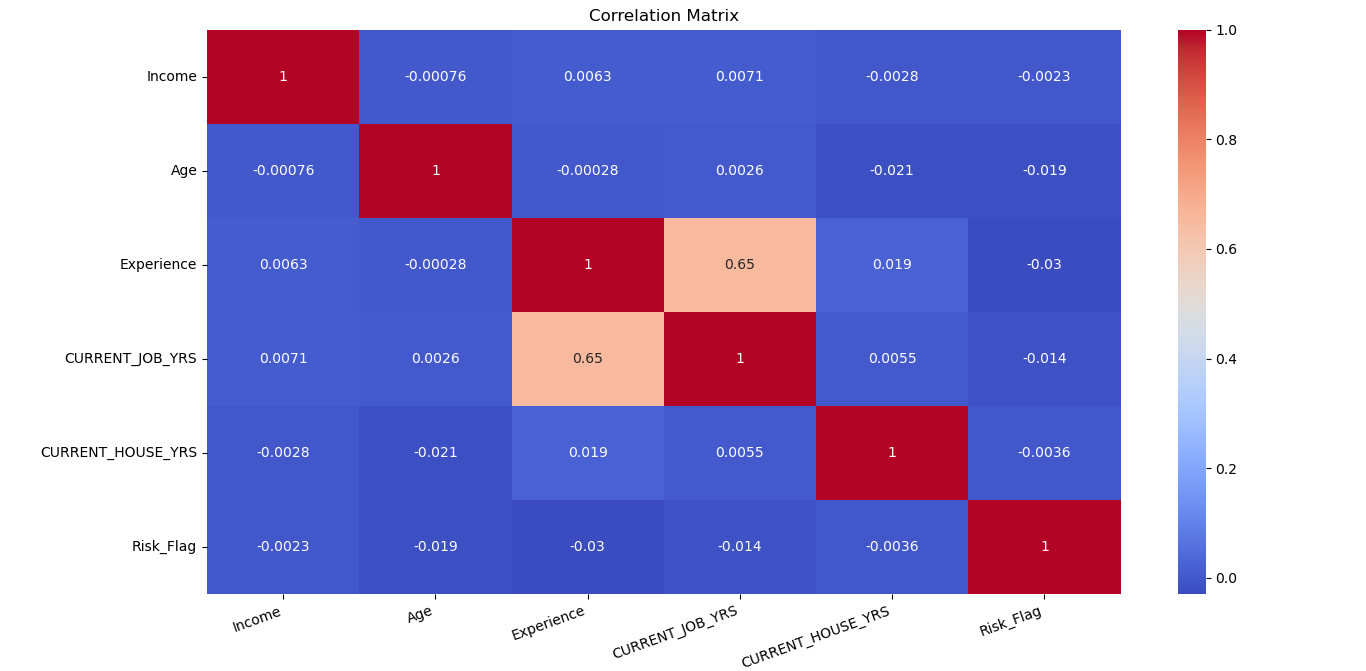
**EDA**

\*کدهای مربوط به این قسمت در فایل EDA ذخیره شده است.

1. ابتدا نمودار هیستوگرام داده های عددی را رسم کردیم.

با توجه به نمودارهای بالا مقادیر پرت مانند داده های منفی یا بیش از حد بزرگ وجود ندارد. اما متغیر، risk\_flag داده های imbalance دارد که ممکن است باعث آموزش کمتر روی یک گروه و نتایج ضعیف تر روی آن گروه بشود که در این صورت می توان با روش های مختلف مانند over\_sampling یا under\_sampling و یا روش های دیگر این مشکل را حل کرد.

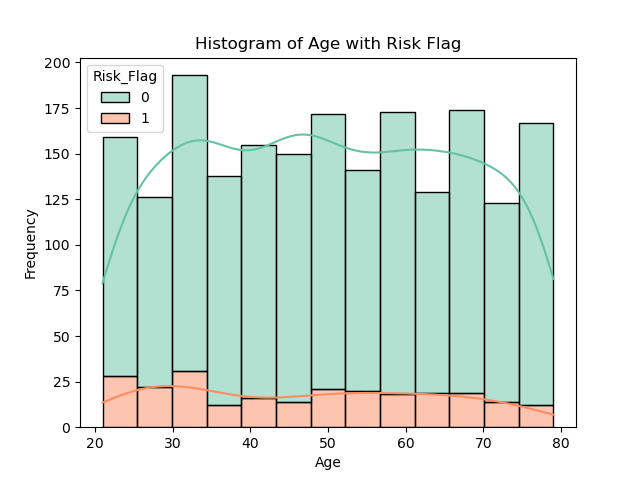
1. در مرحله بعد correlation بین ویژگی ها چک می کنیم.

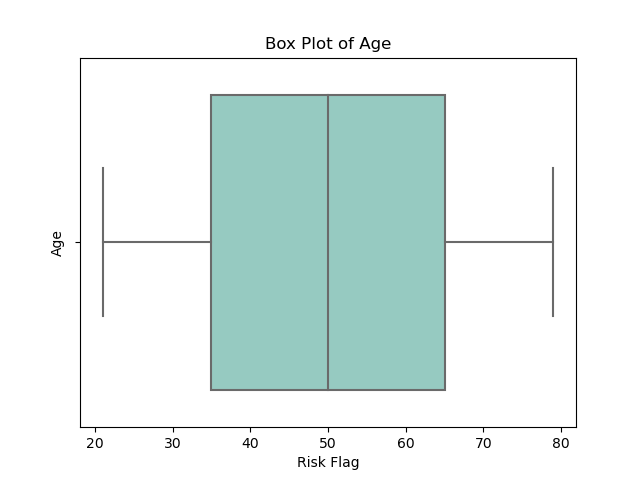


همانطور که مشاهده می شود دو ویژگی experience و current\_job\_yrs همبستگی بالایی دارند و می توان یکی از آن ها را از مجموعه داده ها حذف کرد.

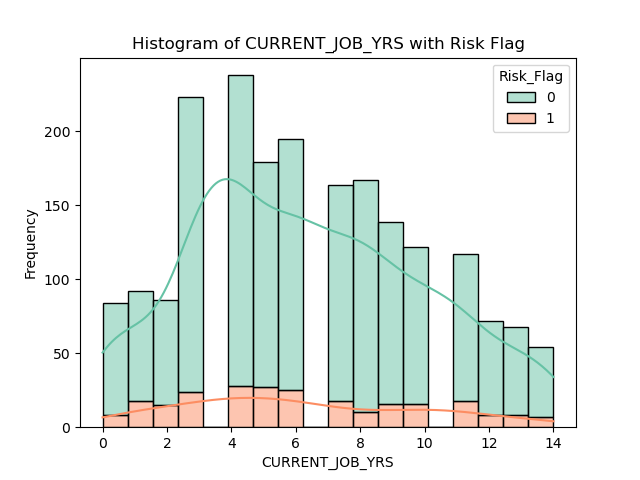
1. هر کدام از ویژگی ها را تحلیل بیشتری می کنیم.

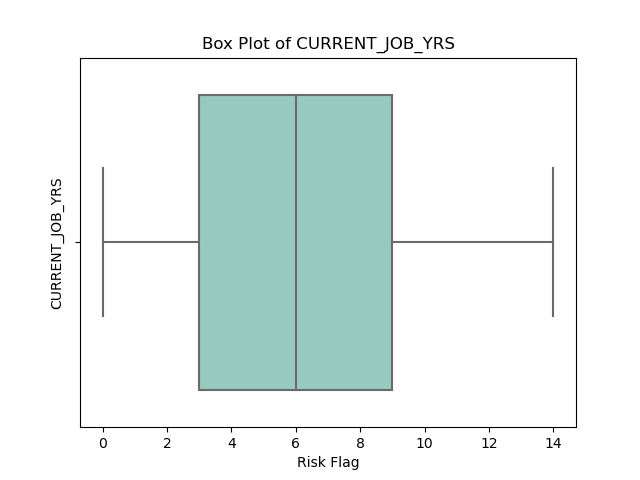
**Age:**

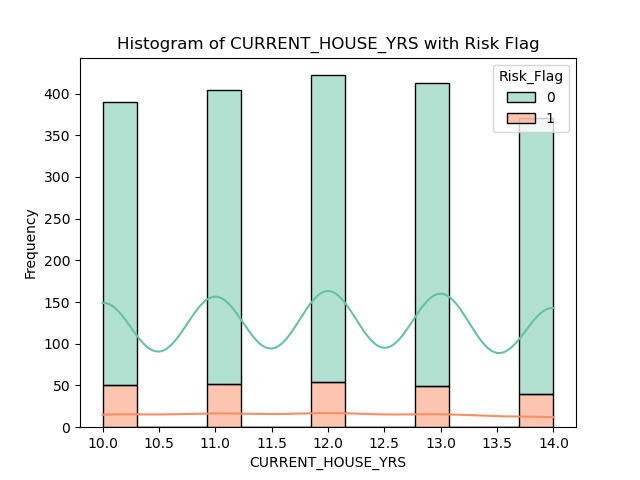


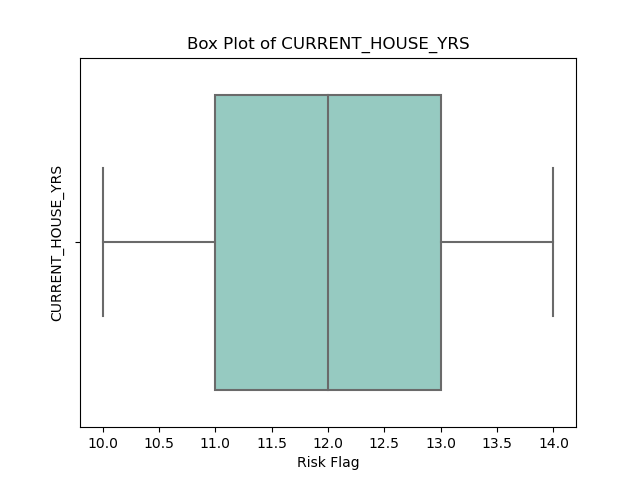


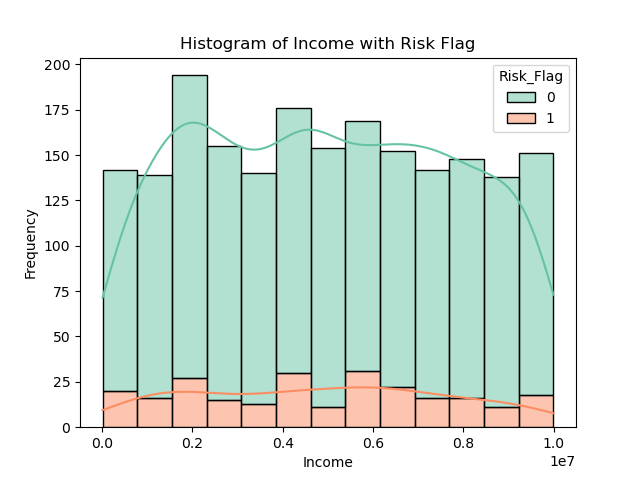
**Current\_job\_yrs:**

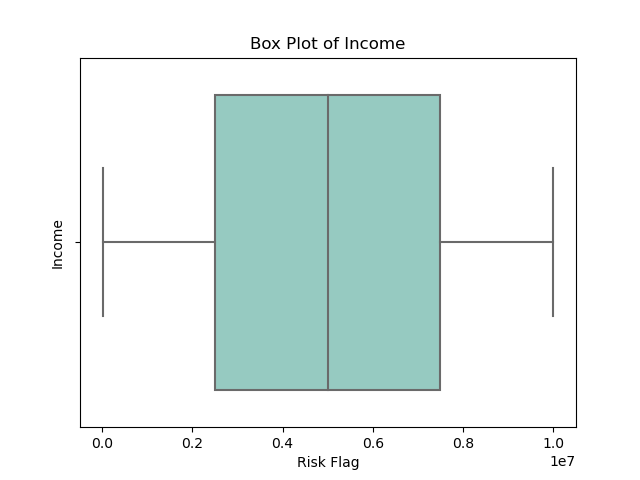




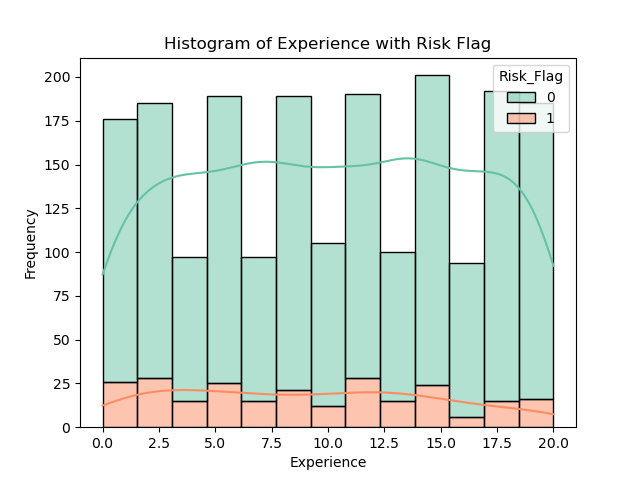
**Current\_house\_yrs:**

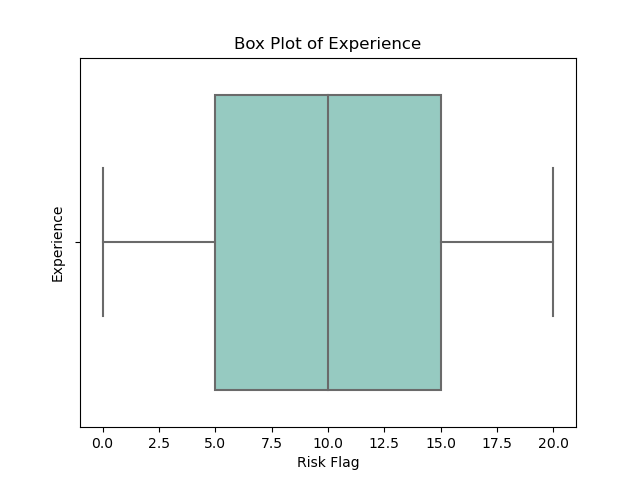


**Income:**

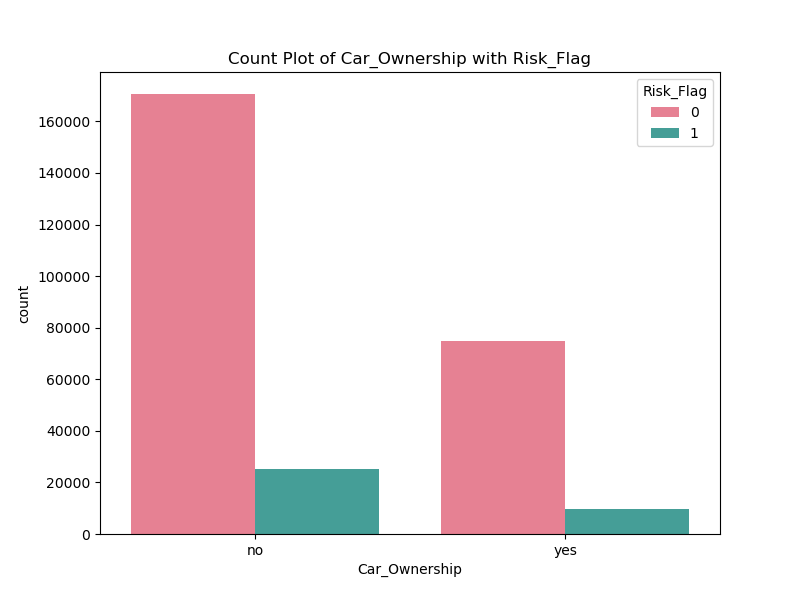


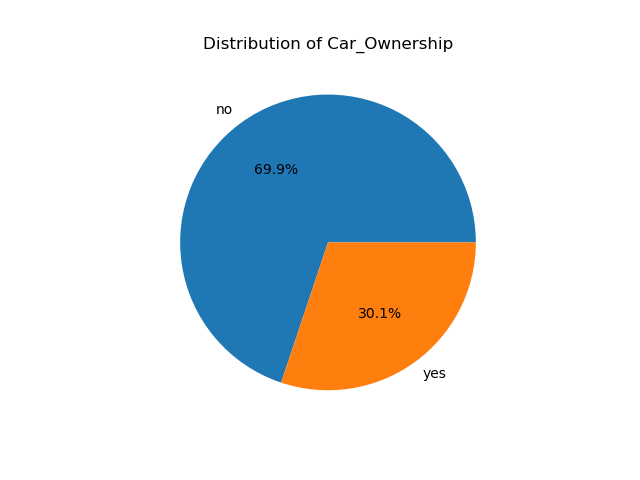
**Experience:**



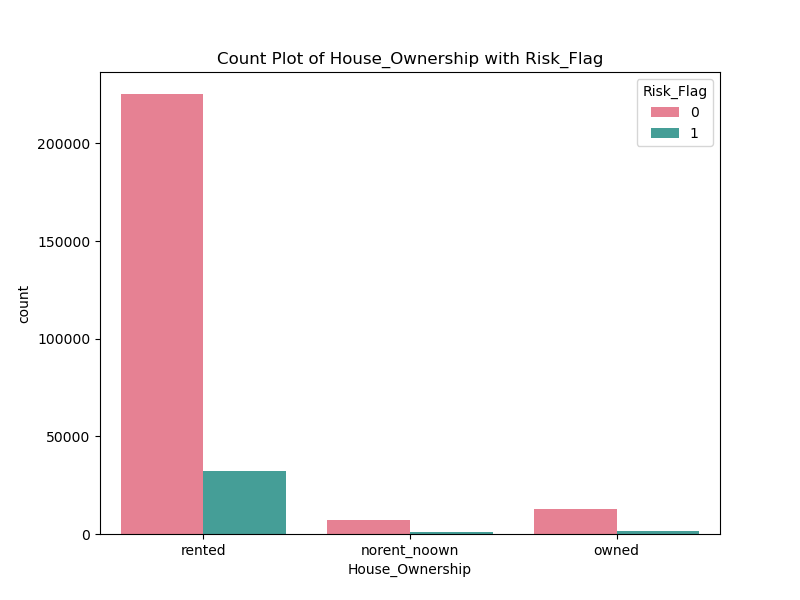


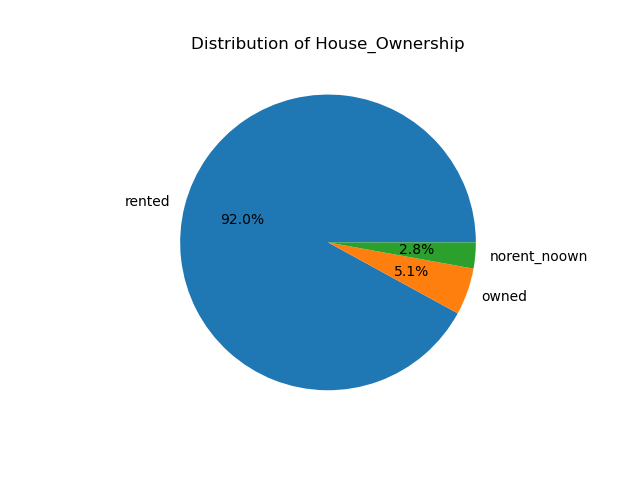
**Car\_ownership:**



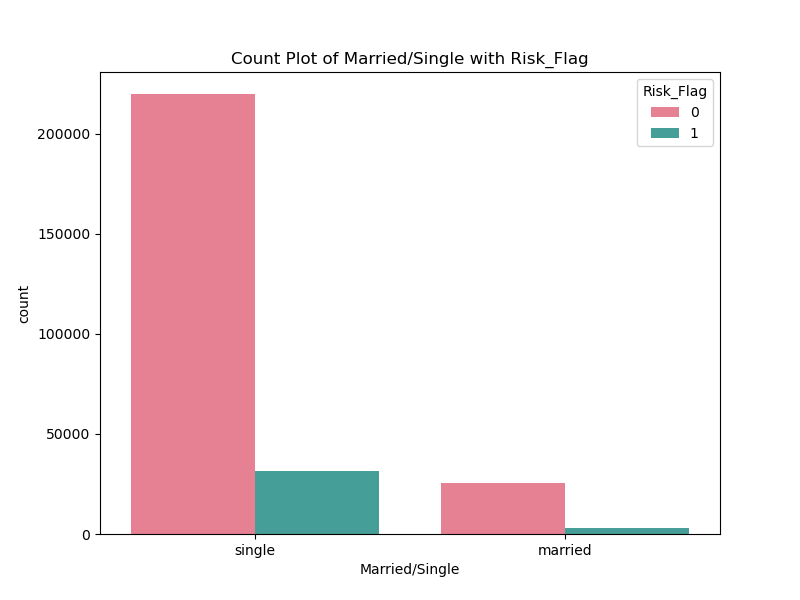


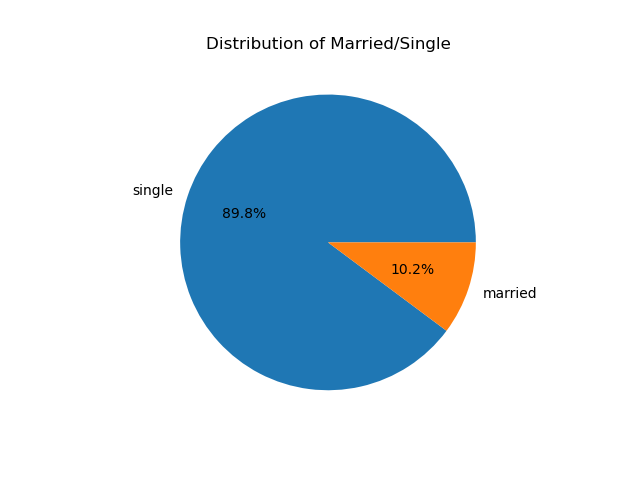
**House\_ownership:**





**Married/single:**

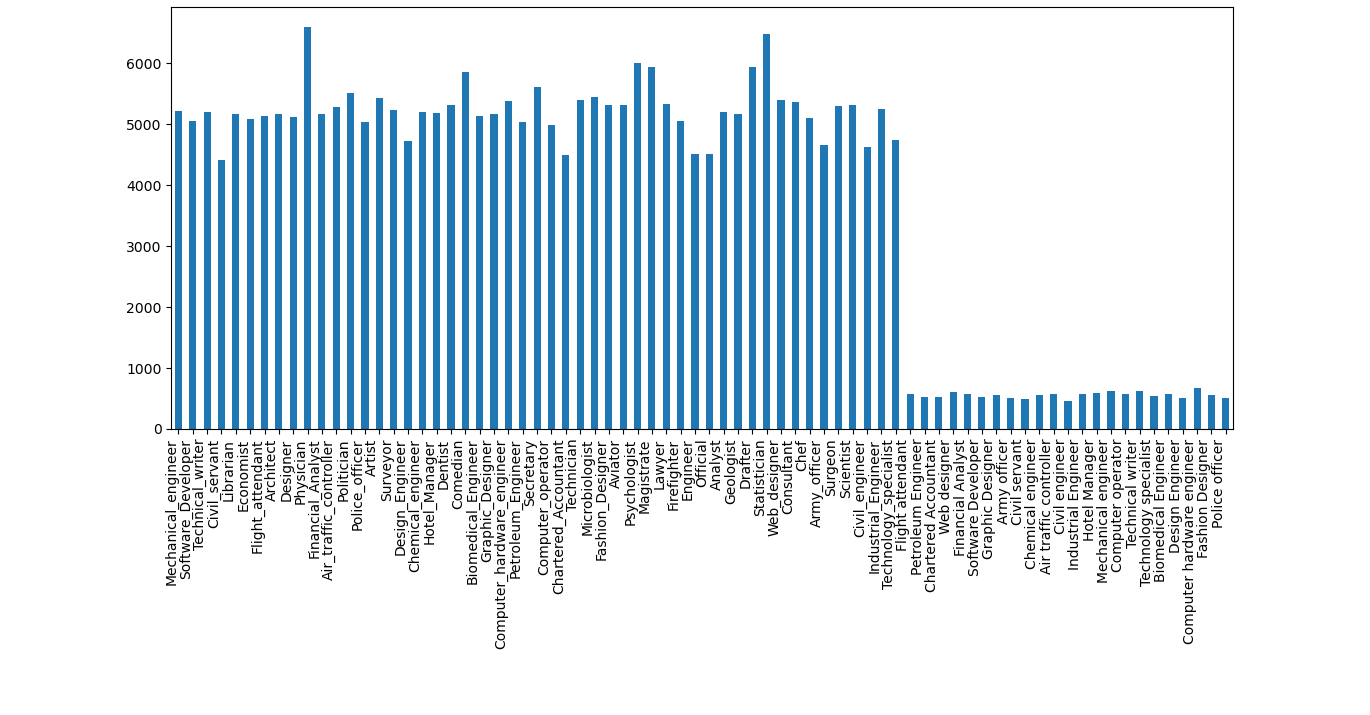


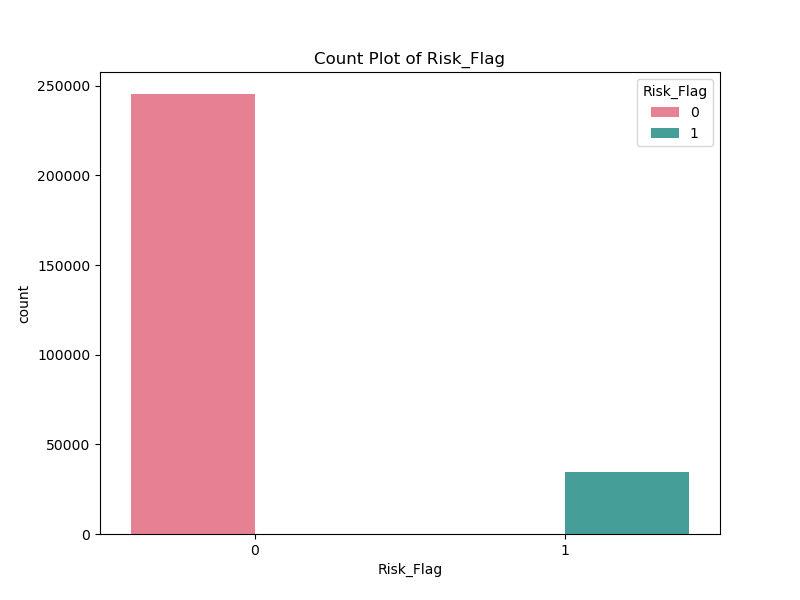


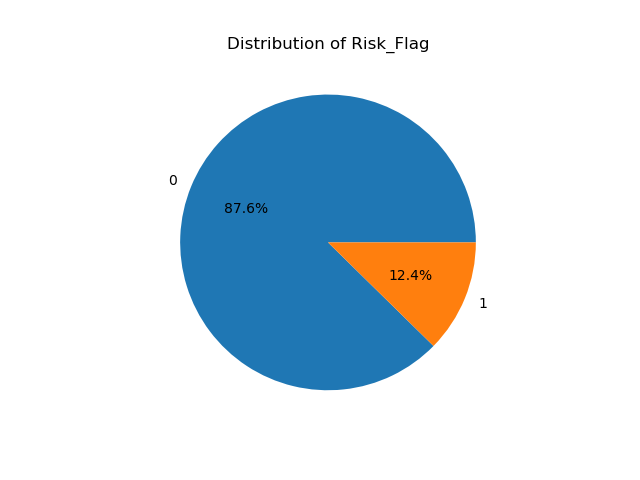
**Professional:**

آیتم های ویژگی Professional به صورت زیر است. همانطور که مشاهده می شود شغل های مشابه بسیار زیاد است که فقط در املایشان با هم متفاوت هستند. همچنین با توجه به تعداد بالای آیتم ها نمودارهای حاصله مناسب برای تفسیر نیستند.

Mechanical\_engineer, Software\_Developer, Technical\_writer, Civil\_servant, Librarian, Economist, Flight\_attendant, Architect, Designer, Physician, Financial\_Analyst, Air\_traffic\_controller, Politician, Police\_officer, Artist' 'Surveyor, Design\_Engineer, Chemical\_engineer, Hotel\_Manager, Dentist, Comedian, Biomedical\_Engineer, Graphic\_Designer, Computer\_hardware\_engineer, Petroleum\_Engineer, Secretary, Computer\_operator, Chartered\_Accountant, Technician, Microbiologist, Fashion\_Designer, Aviator, Psychologist , Magistrate, Lawyer, Firefighter, Engineer, Official, Analyst, Geologist, Drafter, Statistician, Web\_designer, Consultant, Chef, Army\_officer, Surgeon, Scientist, Civil\_engineer, Industrial\_Engineer, Technology\_specialist, Flight attendant, Petroleum Engineer, Chartered Accountant, Web designer, Financial Analyst, Software Developer, Graphic Designer, Army officer, Civil servant, Chemical engineer, Air traffic controller, Civil engineer, Industrial Engineer, Hotel Manager, Mechanical engineer, Computer operator, Technical writer, Technology specialist, Biomedical Engineer, Design Engineer, Computer hardware engineer, Fashion Designer, Police officer



**risk flag:**



**City:**

آیتم های موجود در این ویژگی به شرح زیر است:

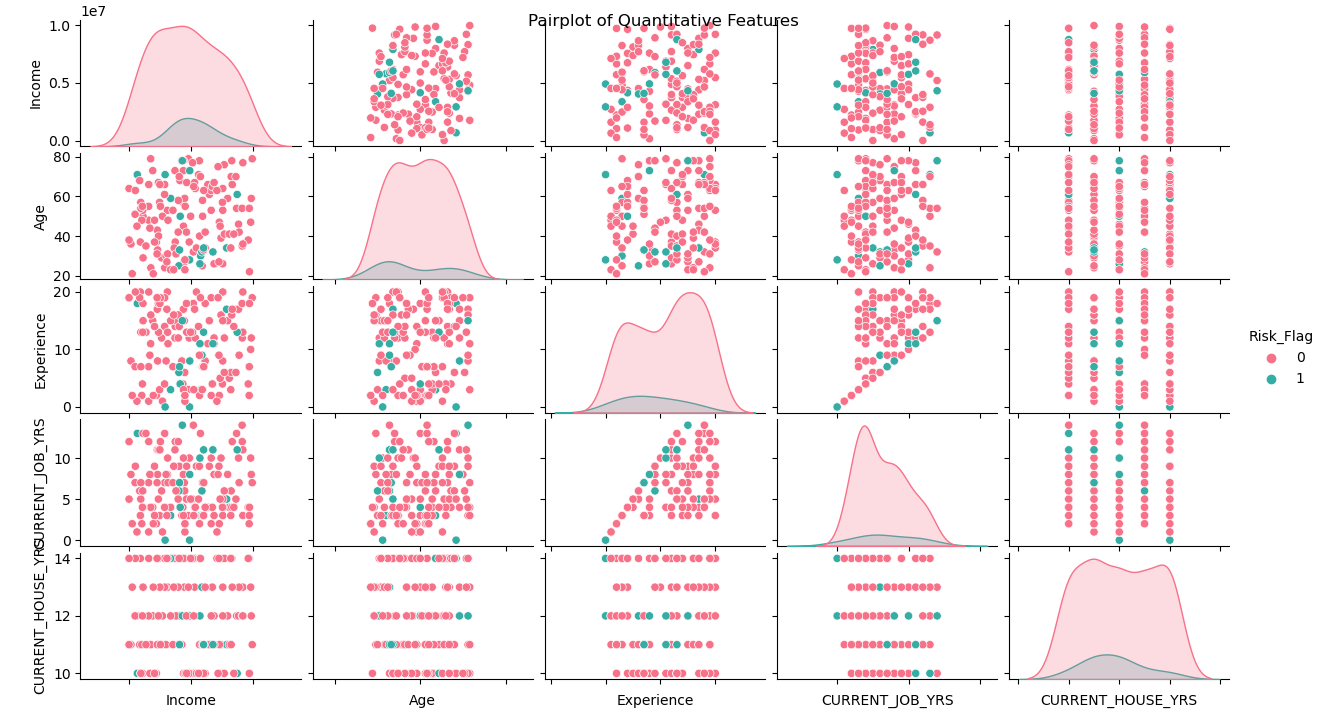
Rewa, Parbhani, Alappuzha, Bhubaneswar, Tiruchirappalli[10], Jalgaon, Tiruppur, Jamnagar, Kota[6], Karimnagar, Hajipur[31], Adoni, Erode[17], Kollam, Madurai, Anantapuram[24], Kamarhati, Bhusawal, Sirsa, Amaravati, Secunderabad, Ahmedabad, Ajmer, Ongole, Miryalaguda, Ambattur, Indore, Pondicherry, Shimoga, Chennai, Gulbarga, Khammam, Saharanpur, Gopalpur, Amravati, Udupi, Howrah, Aurangabad[39], Hospet, Shimla, Khandwa, Bidhannagar, Bellary, Danapur, Purnia[26], Bijapur, Patiala, Malda, Sagar, Durgapur, Junagadh, Singrauli, Agartala, Thanjavur, Hindupur, Naihati, North\_Dumdum, Panchkula, Anantapur, Serampore, Bathinda, Nadiad, Kanpur, Haridwar, Berhampur, Jamshedpur, Hyderabad, Bidar, Kottayam, Solapur, Suryapet, Aizawl, Asansol, Deoghar, Eluru[25], Ulhasnagar, Aligarh, South\_Dumdum, Berhampore, Gandhinagar, Sonipat, Muzaffarpur, Raichur, Rajpur\_Sonarpur, Ambarnath, Katihar, Kozhikode, Vellore, Malegaon, Kochi, Nagaon, Nagpur, Srinagar, Davanagere, Bhagalpur, Siwan[32], Meerut, Dindigul, Bhatpara, Ghaziabad, Kulti, Chapra, Dibrugarh, Panihati, Bhiwandi, Morbi, Kalyan-Dombivli, Gorakhpur, Panvel, Siliguri, Bongaigaon, Patna, Ramgarh, Ozhukarai, Mirzapur, Akola, Satna, Motihari[34], Jalna, Jalandhar, Unnao, Karnal, Cuttack, Proddatur, Ichalkaranji, Warangal[11][12], Jhansi, Bulandshahr, Narasaraopet, Chinsurah, Jehanabad[38], Dhanbad, Gudivada, Gandhidham, Raiganj, Kishanganj[35], Varanasi, Belgaum, Tirupati[21][22], Tumkur, Coimbatore, Kurnool[18], Gurgaon, Muzaffarnagar, Aurangabad, Bhavnagar, Arrah, Munger, Tirunelveli, Mumbai, Mango, Nashik, Kadapa[23], Amritsar, Khora, \_Ghaziabad, Ambala, Agra, Ratlam, Surendranagar\_Dudhrej, Delhi\_city, Bhopal, Hapur, Rohtak, Durg, Korba, Bangalore, Shivpuri, Thrissur, Vijayanagaram, Farrukhabad, Nangloi\_Jat, Madanapalle, Thoothukudi, Nagercoil, Gaya, Chandigarh\_city, Jammu[16], Kakinada, Dewas, Bhalswa\_Jahangir\_Pur, Baranagar, Firozabad, Phusro, Allahabad, Guna, Thane, Etawah, Vasai-Virar, Pallavaram, Morena, Ballia, Surat, Burhanpur, Phagwara, Mau, Mangalore, Alwar, Mahbubnagar, Maheshtala, Hazaribagh, Bihar\_Sharif, Faridabad, Lucknow, Tenali, Barasat, Amroha, Giridih, Begusarai, Medininagar, Rajahmundry[19][20], Saharsa[29], New\_Delhi, Bhilai, Moradabad, Machilipatnam, Mira-Bhayandar, Pali, Navi\_Mumbai, Mehsana, Imphal, Kolkata, Sambalpur, Ujjain, Madhyamgram, Jabalpur, Jamalpur[36], Ludhiana, Bareilly, Gangtok, Anand, Dehradun, Pune, Satara, Srikakulam, Raipur, Jodhpur, Darbhanga, Nizamabad, Nandyal, Dehri[30], Jorhat, Ranchi, Kumbakonam, Guntakal, Haldia, Loni, Pimpri-Chinchwad, Rajkot, Nanded, Noida, Kirari\_Suleman\_Nagar, Jaunpur, Bilaspur, Sambhal, Dhule, Rourkela, Thiruvananthapuram, Dharmavaram, Nellore[14][15], Visakhapatnam[4], Karawal\_Nagar, Jaipur, Avadi, Bhimavaram, Bardhaman, Silchar, Buxar[37], Kavali, Tezpur, Ramagundam[27], Yamunanagar, Sri\_Ganganagar, Sasaram[30], Sikar, Bally, Bhiwani, Rampur, Uluberia, Sangli-Miraj\_&\_Kupwad, Hosur, Bikaner, Shahjahanpur, Sultan\_Pur\_Majra, Vijayawada, Bharatpur, Tadepalligudem, Tinsukia, Salem, Mathura, Guntur[13], Hubliâ€“Dharwad, Guwahati, Chittoor[28], Tiruvottiyur, Vadodara, Ahmednagar, Fatehpur, Bhilwara, Kharagpur, Bettiah[33], Bhind, Bokaro, Karaikudi, Raebareli, Pudukkottai, Udaipur, Mysore[7][8][9], Panipat, Latur, Tadipatri, Bahraich, Orai, Raurkela\_Industrial\_Township, Gwalior, Katni, Chandrapur, Kolhapur, Bihar Sharif, Sri Ganganagar, Ghaziabad, New Delhi, Kirari Suleman Nagar, Sultan Pur Majra, Rajpur Sonarpur, Nangloi Jat, North Dumdum, South Dumdum, Navi Mumbai, Raurkela Industrial Township, Surendranagar Dudhrej, Bhalswa Jahangir Pur, Karawal Nagar, Sangli-Miraj & Kupwad

**State:**

آیتم های موجود در این ویژگی به شرح زیر است:

Madhya\_Pradesh, Maharashtra, Kerala, Odisha, Tamil\_Nadu, Gujarat, Rajasthan, Telangana, Bihar, Andhra\_Pradesh, West\_Bengal, Haryana, Puducherry, Karnataka, Uttar\_Pradesh, Himachal\_Pradesh, Punjab, Tripura, Uttarakhand, Jharkhand, Mizoram, Assam, Jammu\_and\_Kashmir, Delhi, Chhattisgarh, Chandigarh, Uttar\_Pradesh[5], Manipur, Sikkim, West Bengal, Tamil Nadu, Andhra Pradesh, Madhya Pradesh, Uttar Pradesh, Uttar Pradesh[5], Jammu and Kashmir, Himachal Pradesh

1. نمودار pair plot داده ها را برای تفسیر بیشتر توزیع دو به دوی آن ها با هم رسم می کنیم:



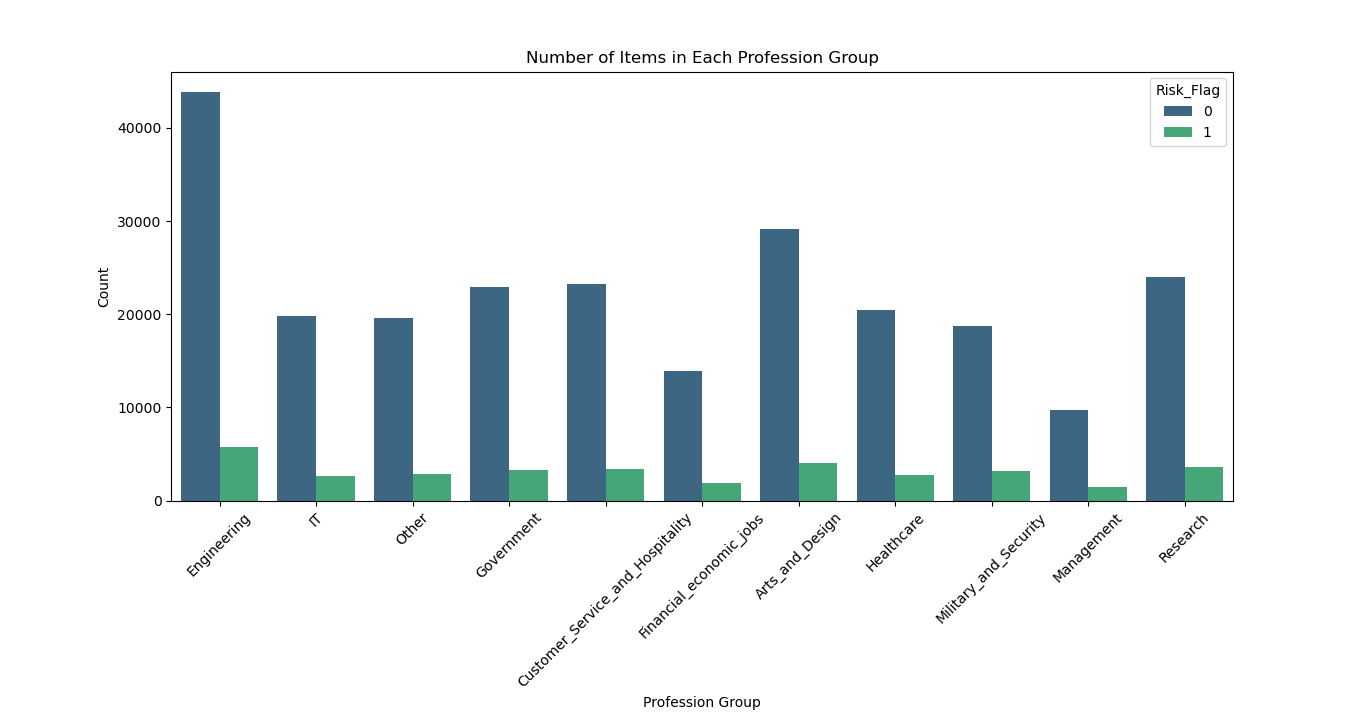
**Feature engineering**

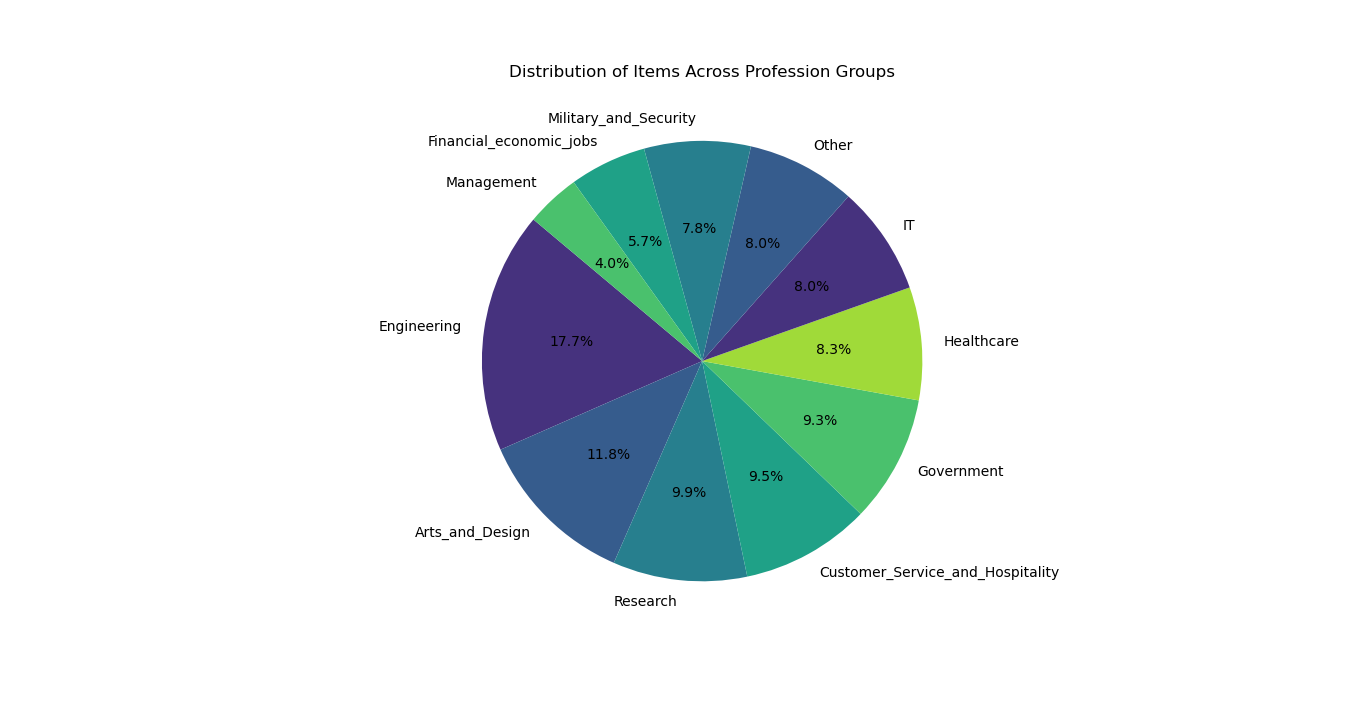
\*کدهای مربوط به این قسمت در فایل Feature engineering ذخیره شده است.

1. ویژگی profession را به گروه های مشابه شغلی به صورت زیر تقسیم کردیم. درنهایت این ویژگی را حذف کرده و ویژگی جدیدی به نام profession\_group ساختیم.

* Engineering Jobs: 'Computer\_hardware\_engineer', 'Industrial\_Engineer', 'Mechanical\_engineer', 'Chemical\_engineer', 'Biomedical\_Engineer', 'Design\_Engineer', 'Civil\_engineer', 'Petroleum\_Engineer', 'Engineer', 'Biomedical Engineer', 'Chemical engineer', 'Civil engineer', 'Computer hardware engineer', 'Design Engineer', 'Industrial Engineer', 'Mechanical engineer', 'Petroleum Engineer'
* IT Jobs: 'Web\_designer', 'Software\_Developer', 'Computer\_operator', 'Technology\_specialist', 'Computer operator', 'Software Developer', 'Technology specialist', 'Web designer'
* Healthcare Jobs: 'Physician', 'Psychologist', 'Dentist', 'Surgeon'
* Management Jobs: 'Hotel\_Manager', 'Consultant', 'Hotel Manager'
* Arts and Design Jobs: 'Fashion\_Designer', 'Graphic\_Designer', 'Artist', 'Designer',
* 'Architect', 'Comedian', 'Fashion Designer', 'Graphic Designer'
* Research Jobs: 'Statistician', 'Microbiologist', 'Scientist', 'Analyst', 'Geologist'
* Government Jobs: 'Lawyer', 'Politician', 'Civil\_servant', 'Official', 'Magistrate', 'Civil servant'
* Military and Security Jobs: 'Air\_traffic\_controller', 'Army\_officer', 'Police\_officer', 'Aviator', 'Air traffic controller', 'Army officer', 'Police officer'
* Customer Service and Hospitality Jobs: 'Flight\_attendant', 'Technician', 'Surveyor', 'Chef', 'Librarian', 'Flight attendant'
* Financial and Economic Jobs: 'Financial\_Analyst', 'Economist', 'Chartered\_Accountant', 'Chartered Accountant', 'Financial Analyst'
* Other: 'Technical\_writer', 'Secretary', 'Firefighter' 'Drafter', 'Technical writer'

نمودارهای ویژگی profession\_group پس از تغییر به صورت زیر است:





* می توان ویژگی سن، تجربه، درآمد و سایر ویژگی های عددی را نیز به صورت های زیر دسته بندی کرد. به عنوان مثال ویژگی سن را می توان به سه گروه افراد جوان، میان سال و پیر تقسیم کرد یا این ویژگی را به بازه های 10 سال 10 سال تقسیم نمود. ویژگی درآمد را نیز به سه یا چهار گروه می توان تقسیم کرد. سال های تجربه را نیز می توان به سه گروه افراد با تجربه زیاد، افراد با تجربه متوسط و افراد با تجربه کم تقسیم کرد. سایر ویژگی ها را نیز به همین ترتیب می توان تبدیل کرد. اگرچه ما برای ابتدای کار با همان ویژگی های عددی ادامه می دهیم و با توجه به نتایج مدل می توان ویژگی ها را تغییر داد.
* اگرچه بهتر است برای دسته بندی تمام ویژگی ها از الگوریتم های clustering مانند kmeans یا hierarchical نیز بهره برد تا در دسته بندی کردن به ما کمک نمایند.
* همچنین ویژگی city و state را که پارامترهای بسیار زیادی داشتند در ابتدای کار حذف کردیم.

**Preprocessing**

\*کدهای مربوط به این قسمت در فایل Preprocessing ذخیره شده است.

1. Scaling

برای scale کردن داده های عددی دو روش min max scaler و standard scaler را امتحان کردیم و در نهایت با توجه به نتایج بهتر روش standard scaler، از آن استفاده کردیم.

1. Encoding

برای کدبندی کردن داده های دسته ای نیز از دو روشone hot encoding و lable encoding استفاده کردیم و در نهایت روش lable encoding نتایج بهتری برای ما داشت. همچنین برای دو ویژگی city و state از الگوریتم count encoder استفاده کردیم و به جای این دو ویژگی، ویژگی های city\_count و state\_count را ساختیم.

\*در نهایت داده ها را به همان صورت اولیه به داده های آموزش و تست تقسیم کردیم و در دو فایل training scaled encoded Data و test scaled encoded Data تقسیم کردیم.

**Modeling**

\*کدهای مربوط به این قسمت در فایل modeling ذخیره شده است.

1. با روش grid search برای الگوریتم های decision tree، xgboost هایپر پارامترهای مدل را بهینه یابی کردیمو نتایج آن به صورت زیر است:

**Decision tree:**

criterion='gini', max\_depth=None, min\_samples\_leaf=4, min\_samples\_split=2

**xgboost**:

learning\_rate=0.3, max\_depth=7, n\_estimators=200

1. با استفاده از cross\_val\_score نتایج برای هر الگوریتم به صورت زیر است:

**Decision tree:**

* **Cross Validation Scores**: [0.91504762 0.89869048 0.83205952]
* **Average Cross Validation Accuracy**: 0.8819325396825396
* **Accuracy on Test Data**: 0.7626785714285714
* **Classification Report on Test Data:**

precision recall f1-score support

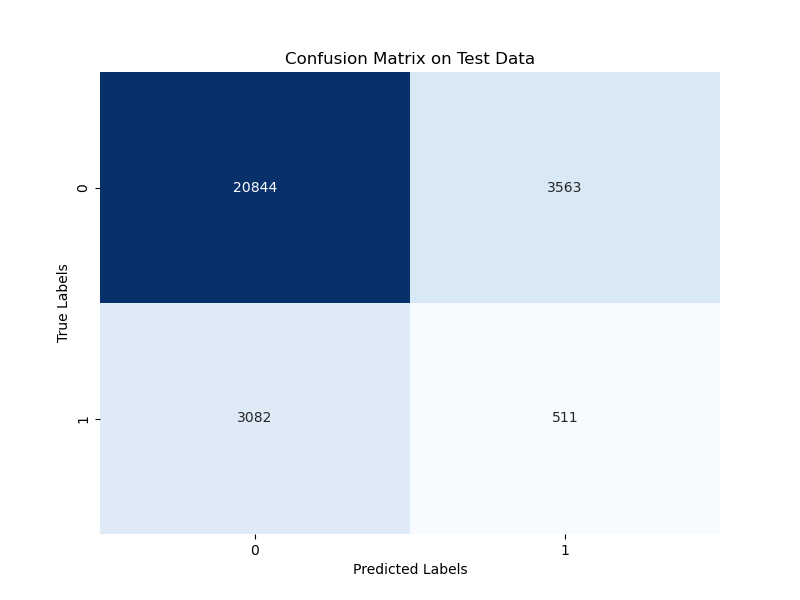
0 0.87 0.85 0.86 24407

1 0.13 0.14 0.13 3593

accuracy 0.76 28000

macro avg 0.50 0.50 0.50 28000

weighted avg 0.78 0.76 0.77 28000



Best

**xgboost**:

**Cross Validation Scores**: [0.90734524 0.89865476 0.87116667]

**Average Cross Validation Accuracy**: 0.8923888888888888

**Accuracy on Test Data**: 0.8198928571428571

**Classification Report on Test Data**:

precision recall f1-score support

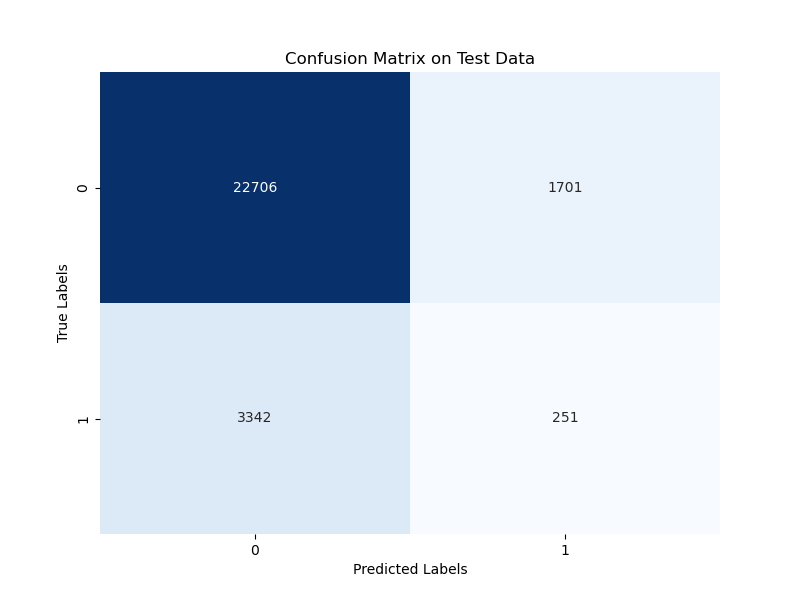
0 0.87 0.93 0.90 24407

1 0.13 0.07 0.09 3593

accuracy 0.82 28000

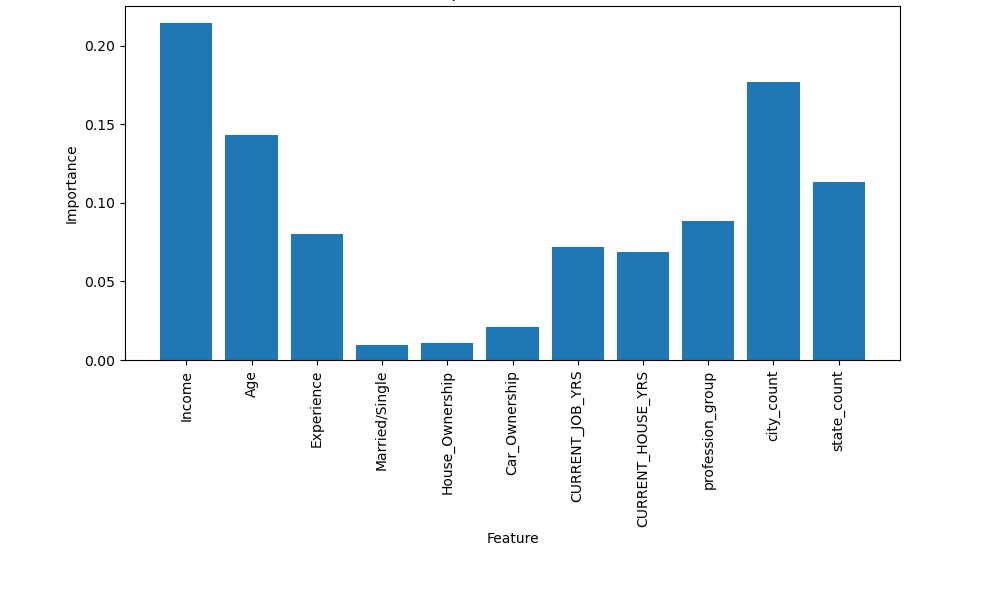
macro avg 0.50 0.50 0.50 28000

weighted avg 0.78 0.82 0.80 28000

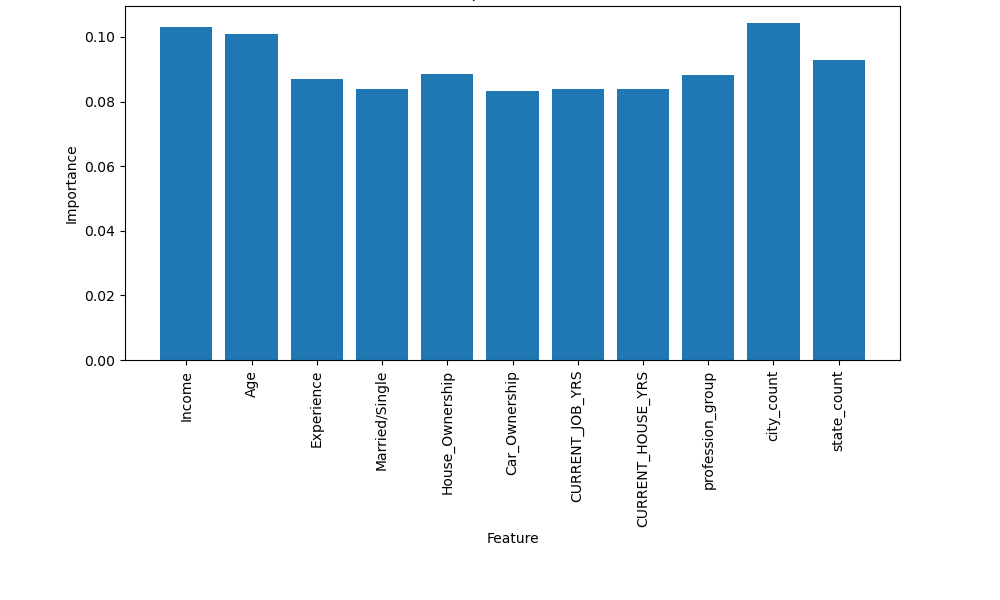


1. Feature importance

مهم ترین ویژگی ها را نیز با دو الگوریتم decision tree و xgboostبه دست آوردیم که نمودارهای آن به صورت زیر است:

Decision tree:

Xgboost:



همانطور که مشاهده می شود در الگوریتم dc اهمیت ویژگی هایhouse\_ownership,carownership, married/single ،کمتر از سایر ویژگی ها است. و ویژگی های Income,city\_count,state\_count,age بیشترین اهمیت را در پیش بینی مدل داشتند. در الگوریتم xgboost نیز نتایج مشابه است با این تفاوت که اهمیت ویژگی های house\_ownership,carownership, married/single با اختلاف کمتری نسبت به سایر ویژگی ها کم اهمیت تر است.

**نتیجه گیری**: با توجه به نتایج به نظر می رسد که مدل over fitشده است و باید بهبود یابد. همچنین با توجه به نتایج مربوط به پیش بینی نکول وام risk flag=1)) و با توجه به این که در قسمت EDA دیدیم که داده ها imbalance هستند می توانیم از روش های over\_sampling و under\_sampling برای ایجاد تعادل در داده ها استفاده کنیم. برای بهبود مدل می توان داده های عددی را به داده های دسته ای تبدیل کرد. برای دسته بندی کردن داده ها نیز می توان از الگوریتم های clustering مانند kmeans یا hierarchical نیز استفاده نمود. همچنین می توان با توجه به اهمیت ویژگی ها، ویژگی های کم اهمیت تر را حذف کرد یا از خود ویژگی های city و state به عنوان یک ویژگی مستقل استفاده کرد و نتایج را دید. همچنین ویژگی هایی که همبستگی بالایی با هم داشتند را نیز می توان برای جلوگیری از پیچیدگی زیاد مدل و جلوگیری از over fit شدن حذف کرد. به طور کلی مقادیر دقت الگوریتم xgboost از dc بهتر بوده است اما برای تعیین مهم بودن این تفاوت دقت می توان از آزمون McNemar’s استفاده کرد